Investigating the gene regulation of sialic acid catabolism in bacteria

Christopher Richard Horne

June 2019

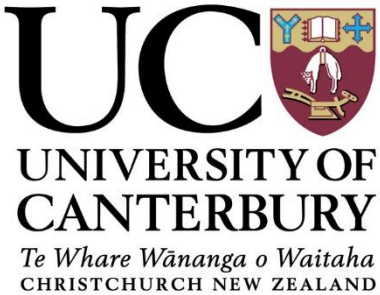
A thesis submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy in Biochemistry

at the University of Canterbury,

School of Biological Sciences



Abstract

Bacteria evolved the ability to uptake and degrade sialic acid as an alternative source of carbon, nitrogen and energy to colonise and persist in environments rich in sialic acid within the human host. The expression of the genes that are responsible for utilisation of this amino sugar are controlled by the transcriptional regulator NanR. The interplay between gene regulation and metabolism of sialic acid is important for efficient growth of the bacterium and to establish infection within the human host. The gap in our knowledge is that the mechanism of this gene regulation is poorly understood at the molecular level. To address this gap, I studied two types of sialic acid gene regulators, which are the RpiR-type NanR from the Gram-positive pathogen *Streptococcus pneumoniae* and the GntR-type NanR from the Gram-negative pathogen *Escherichia coli*.

In the first body of work, I explored the uncharacterised *yjhBC* operon from *E. coli*, which encodes two proteins that are transcriptionally regulated by the GntR-type NanR and that is hypothesised to function in the uptake and processing of less common forms of sialic acid. To test this hypothesis, a thorough *in silico* characterisation suggested YjhB is a sugar-ion symporter within the major facilitator family and identified amino acid residues that may be of functional importance. Using a protein-GFP fusion system, I optimised an overexpression protocol for YjhB, which will be useful for purifying the membrane protein in future biophysical studies. I demonstrated that YjhC is broadly involved in carbohydrate metabolism, through *in vivo* knockout studies. To understand this function further, I solved the crystal structure of YjhC to 1.35 Å, in complex with the cofactor nicotinamide adenine dinucleotide. This verifies the role of YjhC as an oxidoreductase/dehydrogenase, within the Gfo/Idh/MocA family. Using differential scanning fluorimetry and *in silico* docking experiments, which are guided by the structure, I identified several promising leads that suggest YjhC is involved in sialic acid catabolism.

In the second body of work, I found that RpiR-type NanR from *S. pneumoniae* adopts an extended dimeric assembly in solution. This conformation changes to bind DNA, forming a dimeric protein-DNA hetero-complex with a 2:1 binding stoichiometry, as demonstrated by sedimentation velocity and small angle X-ray scattering experiments. This provides the first molecular understanding of the stoichiometry for the RpiR-type NanR-DNA hetero-complex. Multiphase *ab initio* modelling suggested this protein-DNA interaction occurs between the N-terminal DNA-binding domain of NanR to form a binding cleft. To understand the DNA-binding affinity of this interaction, I conducted a kinetic analysis, which revealed a low micromolar affinity. I also demonstrated that the sialic acid pathway metabolite, *N*-acetylmannosamine-6-phosphate binds *S. pneumoniae* NanR and completely abolishes DNA binding. This

supports the hypothesis that *N*-acetylmannosamine-6-phosphate is the effector that attenuates gene repression.

In the final body of work, I demonstrated that the GntR-type NanR from *E. coli* binds as a dimer to a total of three direct GGTATA repeats that make up the DNA recognition site, forming a hexameric protein-DNA hetero-complex, with a 6:1 binding stoichiometry. Interestingly, this interaction is cooperative, mediated by a unique 32 residue N-terminal extension, likely through protein-protein interactions. To understand how DNA binding is attenuated by the effector, I solved the structure of *E. coli* NanR to 2.1 Å in the presence of *N*-acetylneuraminic acid and in complex with a zinc ion. This supports the hypothesis that *N*-acetylneuraminic acid is an effector molecule and identifies zinc as an additional effector. The structure is asymmetrical, with *N*-acetylneuraminic acid and zinc only present in one monomer, informing the molecular choreography that occurs following binding of the effector that attenuates DNA-binding affinity. Notably, the structure of the NanR dimer and DNA hetero-complex was refined to 3.9 Å using cryo-electron microscopy, which is the first known structure of a NanR gene regulator in complex with DNA. In addition, these experiments provided structural evidence of the multimeric assembly process to support the stoichiometry observed in solution. This multimeric protein-DNA assembly is unique to *E. coli* NanR, among known members of the GntR superfamily.

Importantly, this body of work gives the first molecular insight into the mechanism of the NanR-DNA interaction for both a RpiR-type and GntR-type transcriptional regulator, while additionally providing the first structural and functional investigation for both the YjhB and YjhC proteins. Together, this new knowledge enhances our understanding of sialic acid gene regulation and improves our overall grasp of sialic acid catabolism in bacteria.

Deputy Vice-Chancellor's Office
Postgraduate Research Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter Two – Manuscript – 'On the structure and function of *Escherichia coli* YjhC: an oxidoreductase involved in bacterial sialic acid metabolism'

(Submitted to *Proteins: Structure, Function, and Bioinformatics*, 2019)

Chapter Five – Manuscript – 'The cooperative assembly and disassembly of the GntR-type sialoregulator NanR from *Escherichia coli*'

Please detail the nature and extent (%) of contribution by the candidate:

I designed and carried out the experiments, conducted most of the data analysis (90%), and wrote the manuscript.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Christopher Horne* Signature:  Date: *26/06/2019*

Acknowledgements

First and foremost, to my supervisor Professor Renwick Dobson. Thank you for your guidance and for sharing your creative support over the last four years. Thank you for providing me so many amazing opportunities to expand and get the most out of my research. But most importantly, thank you for sharing your incredible passion for science, your unique enthusiasm and always encouraging me to do my best.

A big thankyou to all the members of the Dobson Lab, past and present, for your support and friendship. Special thanks to Jen, Kat and Rachel for your guidance from day one and to the Boyo's (and Serena) of Room 625 for the endless banter or sport related chat. Thanks to Grant Pearce for your insight and wit. Thank you to the extended labs of Level 6, including Callaghan Innovation for your support and sound advice over the years. You are all an incredible bunch of people!

Thank you to Professor Borries Demeler and Deede for hosting me in Canada and providing the guidance to teach me the skills needed to conduct the essential experiments for my project. Thank you to Amy Henrickson for your technical support and for the extended members of the Demeler or Patel Labs for your support while in Canada. Thank you to the MX and SAXS beamline scientists at the Australian Synchrotron for your technical support and guidance over the years. Special thanks to Dr Santosh Panjikar for your assistance in MX and to Dr Hari Venugopal at Monash University for your assistance in Cryo-EM data collection and refinement. Thanks to Dr Janet Newman for your creative assistance.

Thank you to the organisations who have financially supported me over the last four years. The University of Canterbury for a Doctoral Scholarship. The Freemasons NZ for a Postgraduate Scholarship. The Biomolecular Interaction Centre, the Royal Society of New Zealand, and the New Zealand Society for Biochemistry and Molecular Biology for your support to attend conferences in New Zealand and Australia. Thank you to the Maurice Wilkins Centre for funding my research trip to the University of Lethbridge, Canada.

A huge thankyou to my Mum, Dad, Michael, other family and friends outside Uni, for your understanding and support. Thank you for always showing interest in my work, even though you really had no idea what a 'protein crystal' or a 'NanR' was.

Most importantly, thank you to Anna. Thank you for your love, understanding, endless patience and providing encouragement when I needed it most. I could not have completed this journey without you.

Table of contents

Abbreviations

ABC	adenosine triphosphate binding cassette transporter
ADP	adenine diphosphate
ADPS	Adaptive Poisson-Boltzmann Solver
ATP	adenine triphosphate
AUC	analytical ultracentrifuge
BLAST	basic local alignment search tool
C2	carbon position 2 or equivalent
CD	circular dichroism
cm	centimetre
$c(M)$	continuous mass distribution
$c(S)$	continuous sedimentation coefficient distribution
CMP	cytidine monophosphate
Cryo-EM	cryo-electron microscopy
D	diffusion coefficient
Da	dalton
DBD	DNA binding domain
D_{max}	maximum inter-particle dimension
DNA	deoxyribonucleic acid
DSF	differential scanning fluorometry
DTT	dithiothreitol
EBD	effector binding domain
EDTA	ethylenediaminetetraacetic acid
EMSA	electrophoretic mobility shift assay
EMDB	electron microscopy data bank
f/f_0	frictional ratio
F_N	normalised fluorescence
g	gram
GFP	green fluorescent protein
GntR	gluconate (*gnt*) operon transcriptional regulator
Gfo/Idh/MocA	glucose-fructose oxidoreductase/inositol dehydrogenase/rhizopine catabolism family

His-tag	hexanucleotide histidine tag
hr	hour
HTH	helix-turn-helix
I_0	forward scattering value at zero angle
IMAC	immobilised metal affinity chromatography
IPTG	isopropyl β-D-1-thiogalactopyranoside
K_D	dissociation constant
kDa	kilodalton
L	litre
LB	Luria Bertani
LOS	sialylated lipooligosaccharides
LPS	lipopolysaccharides
M	molar
MC	Monte Carlo
MD	molecular dynamics
MFS	major facilitator superfamily
mg	milligram
min	minute
mM	millimolar
mm	millimetre
MRE	mean residue ellipticity
MWL-SV	multi-wavelength sedimentation velocity
NAD	nicotinamide adenine dinucleotide
NADP	nicotinamide adenine dinucleotide phosphate
Nag	*N*-acetylglucosamine catabolic genes
Nan	*N*-acetylneuraminic catabolic genes
NagA	*N*-Acetylglucosamine-6-phosphate deacetylase
NagB	Glucosamine-6-phosphate deaminase
NanA	*N*-Acetylneuraminate lyase
NanE	*N*-Acetylmannosamine-6-phosphate 2-epimerase
NanK	*N*-Acetylmannosamine kinase
NanR	*N*-Acetylneuraminate repressor
Neu5Ac	*N*-Acetylneuraminic acid

nM	nanomolar
nm	nanometer
OD_{600}	optical density at 600 nm
PCR	polymerase chain reaction
PDB	protein data bank
PEG	polyethylene glycol
$P(r)$	real space distance distribution function
R_{factor}	residual factor
R_{free}	free R_{factor}
RFU	relative fluorescence unit
R_{merge}	symmetry ideality
R_g	radius of gyration
RMSD	root mean square deviation
RpiR	ribose-5-phosphate *rpiB* transcriptional regulator
Rpm	revolutions per minute
ROK	repressor, open-reading frame, kinase superfamily
S	sedimentation coefficient (Svedberg)
s	second
SASBDB	small angle scattering biological data bank
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	size exclusion chromatography
SIS	sugar isomerase domain
SSS	sodium solute symporter
SatA	periplasmic binding protein A
SatB/C	integral membrane permease domains
SatD	adenosine triphosphatase domain
SAXS	Small angle X-ray scattering
SiaP	sialic acid binding protein
SiaQ	small transmembrane domain
SiaM	large transmembrane domain
TB	terrific broth
TBE	tris-borate-EDTA
TLS	translation-libration-screw

T_m	melting temperature
TRAP	tripartite adenosine triphosphate independent periplasmic transporter
Tris	2-amino-2-hydroxymethyl-propane-1,3-diol
UDP	uridine diphosphate
v/v	volume to volume
w/v	weight to volume
wHTH	winged helix-turn-helix
2DSA	two-dimensional spectrum analysis
2rDNA	two GGTATA binding site DNA
3D	three-dimensional
3rDNA	three GGTATA binding site DNA
%	percentage
°	degree
°C	degree Celsius
σ	sigma
λ	wavelength
Å	angstrom
µg	microgram
µL	microlitre
µM	micromolar
$\bar{v}$	partial specific volume

Chapter One

Introduction

1.1 Overview

The overall focus of this thesis is the gene regulation of sialic acid catabolism. A broad overview of sialic acid and its importance in bacteria is provided in this introduction. A more detailed review of the literature is presented in the experimental chapters that follow.

1.2 Sialic acid

Sialic acids are a family of nine-carbon amino monosaccharide units (1; 2) that includes more than 50 naturally occurring and structurally distinct variants (3; 4). Structurally, sialic acids contain a carboxylate group, which is deprotonated at physiological pH, resulting in an overall negative charge for the molecule (5). This net negative charge governs the physiochemical properties of sialic acids (1). The arrangement of this carboxylate group (and hydroxyl group) about C2 defines the stereochemistry of sialic acids (Figure 1.1A), existing as an α- or β-anomer (6).

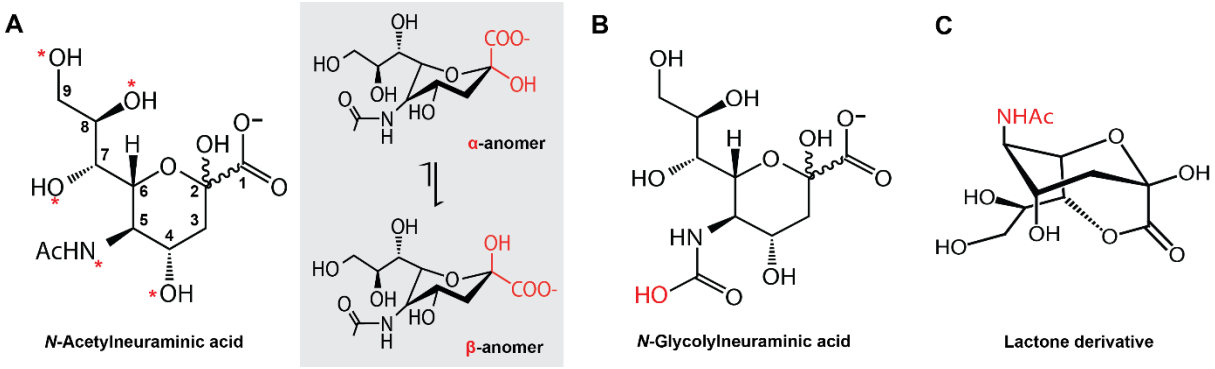


Figure 1.1 Sialic acid. A) *N*-Acetylneuraminic acid (Neu5Ac) is the most abundant and widely studied of the sialic acids. Carbons are labelled one through to nine. The *N*-acetyl group (C5) and the C4, C7, C8 and C9 hydroxyl groups that are subject to substitution/ modification are labelled with a red asterisk. The stereochemical difference at the C2 position defines the α- and β-anomer (grey box). **B)** *N*-Glycolylneuraminic acid is the second most abundant sialic acid, although absent in humans. **C)** The lactone derivative of sialic acid (at positions C1 and C7).

The most common form of sialic acid across all organisms is *N*-acetylneuraminic acid, herein abbreviated to Neu5Ac (Figure 1.1A) (3; 7; 8). The second most abundant form of sialic acid is *N*-glycolylneuraminic acid (Figure 1.1B). Although, this hydroxylated version of Neu5Ac is not present in humans as the hydroxylase gene required for modification is absent (6). Variants are generated through substitutions at the C5 position and modification of the hydroxyl groups at C4, C7, C8 and C9 by acetylation, methylation, sulfation, or in some circumstances the formation of intramolecular esters (lactones) (1; 3). Unlike other derivatives of sialic acid, the lactone variant, formed between the carboxylate and C7 does not carry a net charge or exhibit any defined stereochemistry (Figure 1.1C).

1.3 Significance of sialic acid in eukaryotes

Mammalian cell surfaces are decorated with a complex array of glycoconjugates (9; 10). Located at the termini of many cell surface glycoconjugates are sialic acids (Figure 1.2), where they can mediate cellular interactions, recognition and adhesion processes through immunoglobulin-like lectins, called 'Siglecs' which bind sialic acid (1; 6; 11-13). These processes are fundamental to the pathophysiology of the eukaryotic cells and can act as a defence against infection (2; 13). Overall, the significance of sialic acids in eukaryotes is driven by their distinct chemical diversity (13).

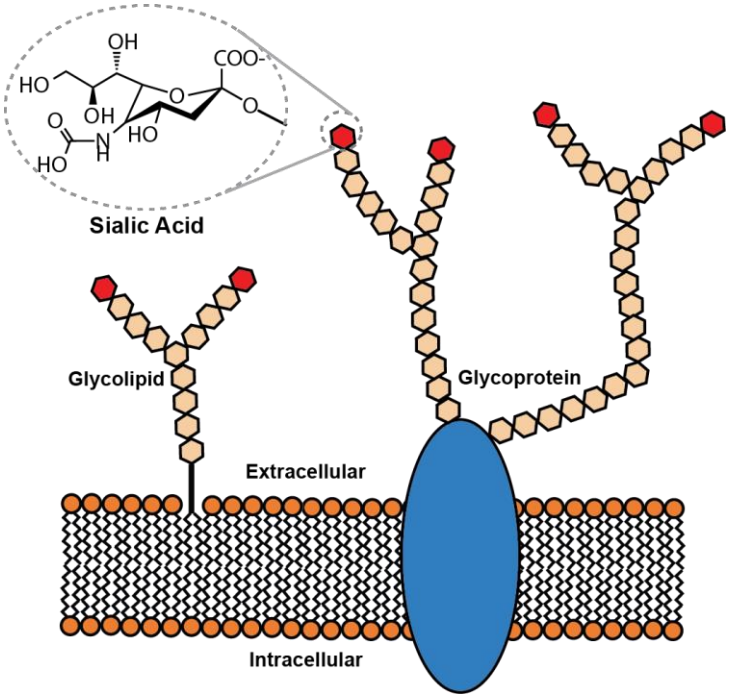


Figure 1.2 | Sialic acids are the terminal sugar on glycoconjugates that coat the surfaces of mammalian cells. Shown is a representation of complex glycoconjugates bound to the surface of the cell, forming a glycolipid and a glycoprotein. Sialic acid (red and insert) is the terminal sugar of these complex glycoconjugates, while sub-terminal sugars (beige) can include *N*-acetylglucosamine, *N*-acetylgalactosamine, glucosamine, galactose, mannose, and fucose (14).

Within the human host, sialic acids are most abundant in the brain where they are utilised in the central nervous system. Their presence in the brain was first discovered by Ernest Klenk over 70 years ago and accounts for the origin of 'neuraminic' (15). The name 'sialic acid' is derived from the Greek word *sialon*, meaning saliva, following the discovery of sialic acid in salivary mucin by Gunnar Blix in a similar era (15). Mucin is a high molecular weight glycoconjugate, present within the mucosal epithelia of the respiratory and gastrointestinal tract (16). Over 65% of terminal glycans are known to contain sialic acid (17; 18) and are believed to provide proteolytic protection to underlying cell surface machinery (19) or serve as the first interaction point for bacteria (9; 10).

1.4 Significance of sialic acid in bacteria

While rich in sialic acid, the respiratory and gastrointestinal tract in humans tend to be glucose-limited and therefore presents an inopportune environment (20; 21). To overcome this limited nutrient availability, bacteria evolved the ability to scavenge sialic acids produced by the human host (7), which are a source of carbon, nitrogen and energy (22). Utilising sialic acid from the human host offers bacteria, such as *Escherichia coli*, a competitive advantage to colonise and persist within the host (23; 24).

In order to utilise sialic acid, bacteria rely on the actions of neuraminidases (25), which hydrolyse the glycosidic linkage of terminal sialic acid residues from glycoconjugates and release them into the surrounding environment. This allows bacteria to scavenge free sialic acid (26; 27). Neuraminidases that release sialic acid are produced endogenously by the host during inflammation (28; 29) or exogenously by neuraminidase-expressing bacteria, such as *Streptococcus pneumoniae* (30) and *Pasteurella multocida* (31). The role of these neuraminidases is essential, since the concentration of sialic acid in the human host within these heavily sialylated niches is relatively high (~2 mM), the majority is inaccessible, locked within complex glycoconjugates (32).

Once the terminal sialic acid is hydrolysed, bacteria can now utilise host-derived sialic acid in a variety of different ways. Primarily, once imported into the cell, sialic acid is catabolised as an alternative nutrient, providing a source of carbon, nitrogen and energy (7; 33). This catabolism of sialic acids also yields precursors for peptidoglycan biosynthesis, such as *N*-acetylglucosamine (34; 35), which is essential for bacterial growth. In addition to sialic acid catabolism, pathogenic bacteria can also decorate their own cell surfaces with host-derived sialic acids. This strategy is known as molecular mimicry and serves as a camouflage mechanism against the host innate immune system (6; 26; 36). Thus, the utilisation of sialic acid in pathogenic bacteria can serve two purposes, by providing a source of nutrition

and playing an important role in the establishment of infection. As a result, sialic acid can act as a virulence factor and play an integral role in pathogenesis (37; 38).

1.5 Uptake of sialic acid by bacteria

The transport of sialic acid into the bacterial cell is an essential process and is required to utilise sialic acid from the surrounding environment (6). This is understood following the discovery that sialic acid uptake is dependent upon a functional transporter in the bacterial pathogens *E. coli* (36), *Haemophilus influenzae* (39) and *Salmonella enterica* (40). Therefore, sialic acid transporters present an attractive target for the development of antimicrobial development towards these pathogenic bacteria.

Once imported, pathogenic bacteria are required to make a choice: catabolise sialic acid as a source of nutrition, or sialylate their cell surface to evade the host innate immune system. Exercising a fine balance between these choices is essential for bacteria to survive and persist within the human host.

The uptake of sialic acid in bacteria is facilitated by multiple transporter families, which vary between different species. These bacterial transporters can be organised into four major types: Secondary proton symporters (36; 41); sodium solute symporters (40; 42); ATP-binding cassette transporters (43); and tripartite ATP-independent periplasmic transporters (39; 44-46).

Generally, the transport of sialic acid is driven by using an energy source or coupling the uptake of sialic acid with the movement of an ion down an electrochemical gradient (47). However, Gram-negative bacteria pose an additional barrier for the uptake of sialic acid with their outer cell membrane. Here, import is governed by the specific porin NanC (48; 49), a bi-partite outer membrane protein complex NanOU (50; 51) and the non-specific porins OmpF and OmpC (52). Collectively, the sialic acid transporters in the outer and inner membrane are discussed below, while their diversity is illustrated in Figure 1.3.

In the bacterial pathogens *E. coli* (36), *Haemophilus influenzae* (39) and *Salmonella enterica* (40), the uptake of sialic acid is dependent upon a functional transporter. This discovery demonstrates that these pathogens utilise a dedicated transporter, specific for the import of sialic acid across the inner plasma membrane, reinforcing this specific transport is an essential process.

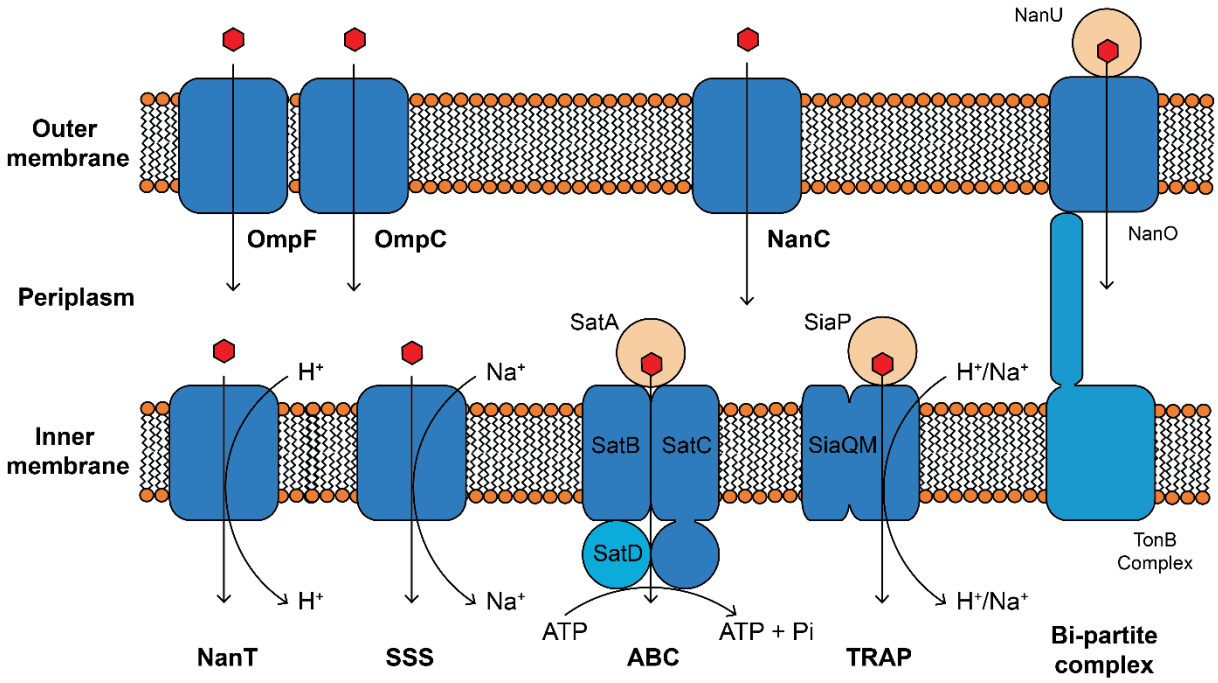


Figure 1.3 | Overview of sialic acid uptake in bacteria. Located in the outer membrane of Gram-negative bacteria, the influx of sialic acid (depicted as a red hexagon) is driven by the specific porin NanC, NanOU, part of the bi-partite complex and the non-specific porins OmpF and OmpC. Located in the inner membrane are four sialic acid-specific transport systems: NanT and the sodium solute symporter (SSS) transporters both consist of a single transmembrane spanning domain and couple the uphill movement of sialic acid with the movement of an ion down its electrochemical gradient; ATP-binding cassette (ABC) transporters consist of a sialic acid-binding protein (SatA), two transmembrane spanning domains (SatBC) and two cytoplasmic nucleotide-binding domains responsible for binding and then hydrolysing ATP to drive the translocation of sialic acid (SatD). In some ABC transporters SatC can be fused with SatD; Tripartite ATP-independent periplasmic (TRAP) transporters consist of a sialic acid-binding protein (SiaP), a small transmembrane spanning domain (SiaQ) and a large membrane spanning domain (SiaM), which can be fused in some TRAP transporters where translocation of sialic acid is driven by an ion motive force.

1.5.1 Uptake of sialic acid across the outer membrane

The outer membrane acts as the first line of defence for Gram-negative bacteria, serving as a semi-permeable barrier to large, charged molecules (53). Therefore, in order to utilise host-derived sialic acid, Gram-negative bacteria must first uptake this resource across the barrier. This influx is controlled specifically by porins through facilitated diffusion or active transport mechanisms (Figure 1.3) (47). The first identified system for the specific uptake of sialic acid across the outer membrane was the *E. coli* NanC porin, encoded by the *nanCMS* operon (49). Although the mechanism is not well understood, the crystal structure identified that the translocation pore is lined by two parallel strings of basic amino acids—this feature is believed to be fundamental for the transport of sialic acid (48).

In *Bacteroides fragilis* and *Tannerella forsythia,* a more complex transport system NanOU, has been discovered, which uses active transport to drive the uptake of sialic acid across the outer membrane. This bi-partite complex is comprised of NanO, a TonB-dependent porin and NanU, an extracellular-binding protein with a high-affinity for sialic acid ($K_D \approx 0.4$ μM). Through structural and functional studies, NanU is believed to first bind sialic acid and then deliver it to NanO before being actively transported into the periplasmic space (50; 51).

Inside the periplasmic space, various processing enzymes have been identified in Gram-negative bacteria to modify sialic acid and mediate its delivery to the inner membrane transporters. Those that modify sialic acid include a sialate mutarotase (*nanM*), which catalyses the periplasmic conversion of the α-anomer of Neu5Ac to the thermodynamically favourable β-anomer (Figure 1.1A) (54), and a 9-O-acetylated Neu5Ac esterase (*nanS*) that permits growth on this derivative of sialic acid (55; 56). Conversely, periplasmic-binding proteins can bind sialic acid with high-affinity and deliver this resource to the ATP-binding cassette and tripartite ATP-independent periplasmic transporters (Section 1.4.4-1.4.5) (43; 57; 58).

In addition to these sialic acid-specific outer membrane transporter systems, influx of sialic acid can also be achieved through simple diffusion, mediated by the general porins, OmpF and OmpC (49). Together, these porins allow an influx of a large variety of molecules and therefore present a broad substrate specificity (52). Within a nutrient limiting environment, these non-specific porins become inefficient. To combat this, pathogenic bacteria evolved these sialic acid-specific transport systems to sustain their survival and persistence within the human host (47).

1.5.2 Sugar-proton symporters

Placed in the major facilitator superfamily, NanT from *E. coli* was the first sialic acid transporter to be discovered (36). Today, NanT is the most common sialic acid transporter amongst Enterobacteriaceae and Bacteroidetes bacterial pathogens (Figure 1.3) (7; 47). Encoded by the *nanT* gene, this bacterial transporter is predicted to be a sugar-proton symporter. To facilitate the transport of sialic acid across the inner membrane, NanT couples the movement of a proton down an electrochemical gradient to facilitate the uphill movement of sialic acid against its electrochemical and/ or concentration gradient in the same direction (59).

The membrane topology of NanT, based on the amino acid sequence is predicted to comprise a total of 14 transmembrane helices. This differs from the traditional 12 transmembrane helices observed among transporters in the major facilitator superfamily. Located centrally, these additional transmembrane

helices (7 and 8) within NanT, are hypothesised to form an amphipathic helix and play a role in the substrate specificity of the transporter (41). The remaining 12 transmembrane helices contain residues that are conserved between members of the major facilitator superfamily (6; 47). Interestingly, a sequence homolog to NanT has been identified in *E. coli*, referred to as YjhB. In Chapter Three, I will investigate the role this transporter may play in sialic acid catabolism.

While the structure of NanT has yet to be solved, the closest structural and biological candidate to NanT with a sequence identity of only 16-23% is the recently reported sugar-proton symporter, XylE from *E. coli*—a D-xylose transporter within the major facilitator superfamily (60). These transporters, among other members of the major facilitator superfamily are proposed to function *via* an alternating access mechanism (59). The N- and C-terminal domains rock back and forth between an inward- and outward-facing open conformation, where the substrate binding site acts as the pivot point only accessible to either the cytoplasmic or periplasmic side of the membrane. This process is known as the 'rocker-switch' mechanism (Figure 1.4) (61; 62).

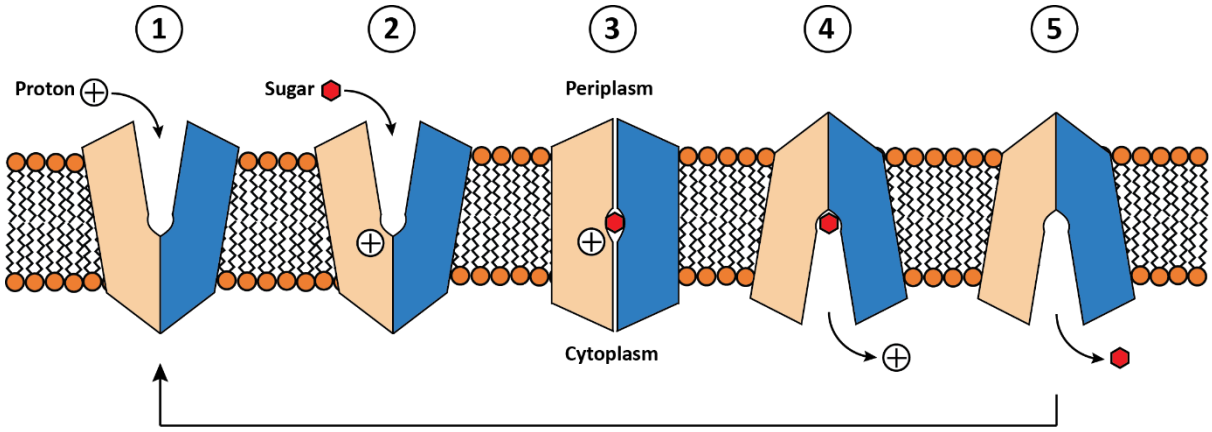


Figure 1.4 | The 'rocker-switch' mechanism. In this mechanism the N-terminal (beige) and C-terminal (blue) domains rock back and forth between an inward- and outward-facing open conformation, where the substrate binding site acts as the pivot point only accessible to either the cytoplasmic or periplasmic side of the membrane. Initially, in an outward-facing open conformation a proton binds from the periplasmic side of the membrane (Position 1). Following the binding of a proton, the substrate binds in a central cavity (Position 2). The binding of the substrate triggers a conformational change, initially through an intermediate occluded state (Position 3), through to an inward-facing open conformation where the proton can exit into the cytoplasm (Position 4). Finally, the substrate is translocated and released into the cytoplasm, completing the cycle (Position 5). While some of the residues involved in both substrate and proton translocation are conserved, others are unique to certain symporters.

1.5.3 Sodium solute symporters

In contrast with sugar-proton symporters, the Sodium Solute Symporter (SSS) family of secondary transporters couple the transport of sodium ions down their electrochemical gradient to facilitate the movement of sugars, amino acids, inorganic ions, or vitamins into the cell (63). This dependency on sodium ions for co-transport is where the family name, SSS originates. Among members of the SSS family, the symporter SiaT has been linked to the genes encoding the processing enzymes for sialic acid, and thus is understood to be a specific transporter for sialic acid.

In 2018, the structure of SiaT from *Proteus mirabilis* was solved (42)—this is the first known structure of a sialic acid transporter. Comprised of 13 transmembrane helices, SiaT adopts a fold analogous to the sodium-leucine symporter, LeuT (42; 64). Interestingly, the structure was solved in the outward-facing, open conformation bound to sialic acid and two sodium ions. While, the binding site for one of these sodium ions is conserved, the other sodium ion binds at a new position and is believed to allosterically stabilise sialic acid. Analogous to the alternating access mechanism (as described for members of the major facilitator superfamily), transport in SiaT also involves the rearrangement of transmembrane helices to translocate sialic acid with the addition of a gating mechanism (42).

1.5.4 ATP-binding cassette transporters

The ATP-binding cassette (ABC) transporters are a ubiquitous superfamily of membrane transporters that couple the energy attained from ATP hydrolysis to transport a variety of different solutes across the outer and inner membranes (65; 66). Within this superfamily, ABC transporters function as both importers and exporters (67). Their typical architecture consists of four domains, which include: two transmembrane domains (forming the translocation pathway), and two cytoplasmic nucleotide-binding domains (responsible for binding and then hydrolysing ATP) (68). These domains can either be fused, which is characteristic of ABC exporters, or can exist as separate domains in ABC importers (65; 69). In addition to this domain architecture, ABC importers can employ substrate-binding proteins to sequester their preferred substrate with high affinity and deliver it to the membrane component (70). These substrate-binding proteins can either be found tethered to the inner membrane in Gram-positive bacteria, or free to scavenge in the periplasmic space (71).

The first report of a bacterial ABC transport system for the uptake of sialic acid was discovered in *Haemophilus ducreyi*, encoded by the genes *satABCD* (43). These genes express a periplasmic-binding protein (SatA), two integral membrane permease domains (SatB/SatC) and an ATPase domain (SatD). Together, these components are named Sat for sialic acid transport. Unique to *H. ducreyi*, the SatC

integral membrane domain is fused with one of the SatD domains (Figure 1.4) (43)—this feature highlights the diversity of the ABC transporter superfamily. Remarkably, the periplasmic-binding protein SatA, is known to demonstrate nanomolar affinity for negatively-charged sialic acid through a process driven by entropy, hydrogen-bonding and electrostatic interactions (72). However, the molecular exchange of sialic acid between SatA and the membrane component, SatBCD is yet to be confirmed. A second class of Sat-type ABC transporters have been identified in several other bacteria, including the pathogen *S. pneumoniae* (73; 74). While this system contains a SatABC domain architecture, it was shown to lack the SatD ATPase domain. Here, ATPase activity is alternatively driven by MsmK, an ATPase responsible for stimulating multiple carbohydrate ABC transporters by interacting with SatABC to energise the translocation of sialic acid (58).

1.5.5 Tripartite ATP-independent periplasmic transporters

The tripartite ATP-independent periplasmic (TRAP) transporters are a unique group of specific solute transporters that only occur in bacteria and archaea and not eukaryotes, making this class of transporters a viable target for antimicrobial development (71). TRAP transporters are known to translocate a range of substrates, including C4-dicarboxylates, α-keto acids, amino acids and various aromatic compounds (75). To drive this translocation, TRAP transporters couple the movement of an ion down an electrochemical gradient to thermodynamically drive the uphill movement of the substrate in the same direction, as opposed to harnessing the energy of ATP (71)—by definition this mechanism is defined as secondary transport.

Structurally, TRAP transporters are composed of three protein domains in a 'PQM' system. These include either a membrane-anchored or periplasmic substrate-binding protein ('P' domain), a large transmembrane spanning domain ('M' domain) consisting of 12 transmembrane helices, believed to form the translocation pore and a small transmembrane spanning domain ('Q' domain) consisting of 4 transmembrane spanning domains (76). This 'Q' domain is poorly conserved and although shown to be essential for transport, its function is largely uncharacterised (46).

The first sialic acid TRAP transporter to be characterised was from the human pathogen, *H. influenzae* (39; 45; 75). Here, the substrate-binding protein (SiaP) delivers sialic acid to the membrane component (SiaQM), where translocation of sialic acid is driven by a sodium motive force (Figure 1.4). Moreover, the SiaQM membrane component is genetically fused in *H. influenzae* (46). Although this feature is not uncommon in TRAP systems, an extra transmembrane domain is required to link the two transmembrane domains—this results in a total of 17 transmembrane helices for *H. influenzae* SiaQM.

Despite identifying a dependency on sodium ions for symport in the H. influenzae TRAP transporter, further characterisation is required to define the ion specificity of sialic acid TRAP transporters (46).

SiaPQM TRAP transporters have also been identified in *P. multocida* (77; 78), *Vibrio vulnificus* (79) and *Vibrio cholerae* (44). Among these, the sialic acid-binding protein, SiaP has been structurally and functionally characterised (80-82). These sialic acid-binding proteins have two domains (Figure 1.5). Between these domains, sialic acid is known to bind with low micromolar to nanomolar affinity within a conserved binding site (72). This binding triggers a conformation change between domains to clamp down on the substrate, mediated through a conserved α-helix hinge region. Comically, this mechanism has been likened to a 'Pac-Man' motion or action of the Venus Fly Trap (83; 84). Like the ABC transporters, the protein-protein interactions and inherent dynamics involved in the exchange of sialic acid between SiaP and SiaQM is yet to be elucidated.

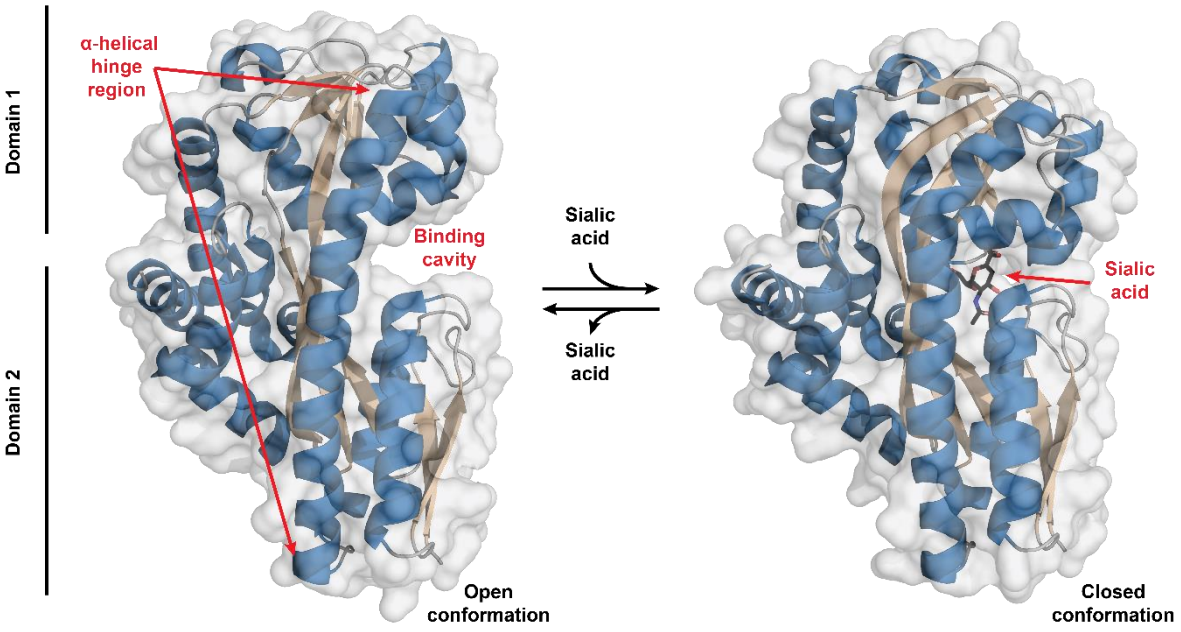


Figure 1.5 | The periplasmic sialic acid-binding protein, SiaP. The left panel presents the structure of the periplasmic sialic acid-binding protein SiaP in its open conformation. Highlighted is the hinge region and binding cavity. The right panel presents the structure of SiaP in its closed conformation, induced through binding of sialic acid—this mechanism has been likened to a 'Pac-Man' motion or action of the Venus Fly Trap (83; 84). To help illustrate this mechanism, the molecular surface has been mapped to each structure, where both open and closed conformations is SiaP from *H. influenzae* (PDB ID 2CEX) (81).

1.6 Cell-surface sialyation in bacteria

Cell-surface sialyation is a prominent feature of eukaryotic cell surfaces (85). Surprisingly, some pathogenic bacteria have evolved the ability to decorate their own cell-surface with sialic acid to mimic the host cell, masquerading themselves in order to evade the host innate immune response (34; 86), and serve as a virulence factor. This process is achieved by incorporating sialic acid into lipopolysaccharides (LPS) of Gram-negative bacteria; sialic acid-containing polysaccharide capsules, or the flagellum (87). The selective advantage of this process is so prominent that at least four separate strategies of cell-surface sialylation have evolved in pathogenic bacteria (6; 34). These strategies include: *de novo* biosynthesis; donor scavenging; a *trans*-neuraminidase mechanism; and precursor scavenging. Collectively, this diversity is illustrated in Figure 1.6.

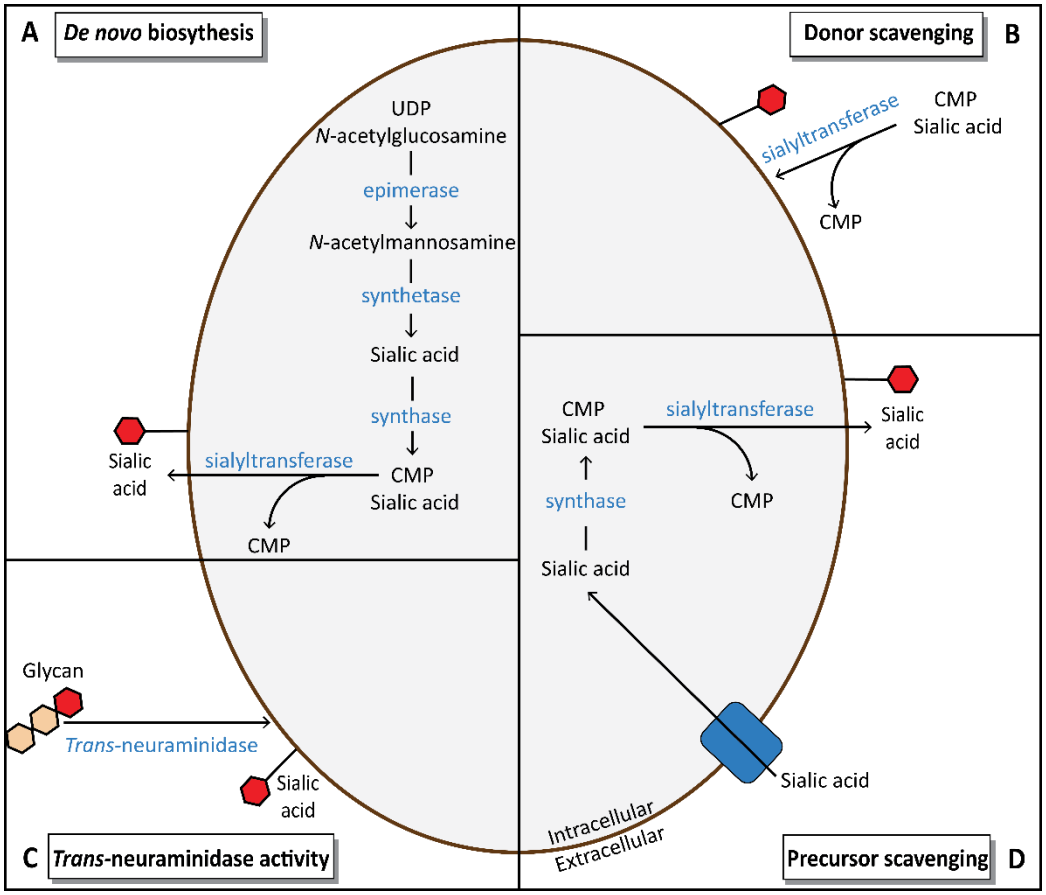


Figure 1.6 | The mechanisms of cell surface sialylation in bacteria. Presented is a bacterial cell (oval) with both intracellular (grey) and extracellular (white) environments. The four mechanisms of cell surface sialylation are: **A)** *De novo* biosynthesis; **B)** Donor scavenging; **C)** *Trans*-neuraminidase activity; and **D)** Precursor scavenging. The specific enzymes and membrane transporter (rectangle) involved in these strategies are coloured blue. Sialic acid is depicted as a red hexagon on the surface of the cell, or at the terminal position of a glycoconjugate (glycan), while sub-terminal sugars are coloured beige.

1.6.1 *De novo* biosynthesis

Several bacteria including *E. coli* K1, *Neisseria meningitidis* and *Campylobacter jejuni* can synthesise their own sialic acid from simple metabolites (Figure 1.6A) (6). This biosynthetic pathway is thought to begin with the enzymatic conversion of the cell wall precursor uridine diphosphate (UDP) *N*-acetylglucosamine, through to sialic acid, *via* several enzymatic steps (88; 89). Once synthesised, sialic acid is converted into cytidine monophosphate (CMP) sialic acid, an activated form of sialic acid, which is subsequently added to appropriate acceptors by linkage-specific sialyltransferases and exported to the cell surface (34). Here, appropriate acceptor molecules include lipopolysaccharides (LPS), sialylated lipooligosaccharides (LOS) and sialic acid-containing polysaccharide capsules (86).

1.6.2 Donor scavenging

In a second strategy, *Neisseria gonorrhoeae* also utilises linkage-specific sialyltransferases and CMP-sialic acid to sialylate their cell surfaces. However, *N. gonorrhoeae* contain a significantly truncated sialylation pathway, deficient in the genes responsible for sialic acid biosynthesis. Therefore, CMP-sialic acid must be scavenged exogenously from the host, rather than synthesised endogenously using UDP *N*-acetylglucosamine (90)—a strategy termed donor scavenging (Figure 1.6B).

1.5.3 *Trans*-neuraminidase activity

This cell surface sialylation strategy identified within *Trypanosoma cruzi* is unique as this organism neither scavenges, nor synthesises sialic acid. Alternatively, *T. cruzi* has a *trans*-neuraminidase mechanism, which has both neuraminidase and sialyltransferase activity (Figure 1.6C) (91). While the neuraminidase activity is responsible for hydrolysing the glycosidic linkage of terminal sialic acid, the sialyltransferase activity is responsible for directly transferring the cleaved sialic acid to the cell surface of the organism, without requiring activated sialic acid (6; 86).

1.6.4 Precursor scavenging

The most recently discovered strategy of cell-surface sialylation, termed precursor scavenging (Figure 1.6D), was identified in the bacterial pathogen *H. influenzae* (92). Like *N. gonorrhoeae, H. influenzae has a* truncated sialylation pathway and is incapable of synthesising sialic acid from the simple metabolite UDP-*N*-acetylglucosamine. As a consequence, it must scavenge host-derived sialic acid by importing directly into its cell (86). Once internalised, the host-derived sialic acid is subsequently converted to CMP-sialic acid, where sialyltransferase incorporates it into LPS, LOS or sialic acid-containing polysaccharide capsules.

1.7 Catabolism of sialic acid by bacteria

Once host-derived sialic acid is made bioavailable to the pathogenic bacteria, enabling transport into the cell, it is then faced with the choice to sialylate their cell-surface and evade the host innate immune system, or utilise the sialic acid as a source of nutrition by degrading it into a supply of carbon, nitrogen and energy (86; 92). This catabolic process is facilitated by six key steps.

The first four steps of the sialic acid catabolic pathway involve a sialic acid-specific transporter, responsible for the uptake of sialic acid into the bacterial cell (Section 1.5), and three catalytic enzymes (22): *N*-acetylneuraminate acid lyase; *N*-acetylmannosamine kinase; and *N*-acetylmannosamine-6-phosphate 2-epimerase—a suite of proteins collectively known as the *N*-acetylneuraminate acid (Nan) cluster. This cluster has been identified in over 46 bacterial species, with the majority represented by human pathogenic bacteria (7).

Following flux through the Nan cluster, the final steps in the catabolic pathway are facilitated by two enzymes, *N*-acetylglucosamine-6-phosphate deacetylase and glucosamine-6-phosphate deaminase. Together, these enzymes are known as the Nag enzymes (*N*-acetylglucosamine).

Although clustered, the distribution of the *nan* genes is variable, where the order of the genes encoding these enzymes is known to vary both among and within bacterial families (6; 7). In comparison, the genes encoding the Nag enzymes are not clustered and are scattered amongst the genome. The only reported exception is within *H. influenzae* where these genes are closely associated with the Nan cluster (6). Taken together, the diversity in sialic acid-specific transporters, the difference in gene organisation and the limited distribution of the essential catabolic genes is indicative of 'mosaic evolution' (6; 7). Put simply, the observed diversity and lack of synteny among bacteria is driven by an individual pressure to adapt to their respective hosts and surrounding environment in order to survive.

Collectively, the metabolic overlap of these enzymes fabricates the core catabolic pathway that functions to degrade sialic acid into fructose-6-phosphate, in order to drive the glycolytic cycle (6), or yield precursors for peptidoglycan biosynthesis, such as *N*-acetylglucosamine (34; 35). An illustrative overview of this catabolic pathway is presented in Figure 1.7, followed by a summary of each key enzyme to emphasise the complexity of this degradative process. Aside from this core catabolic pathway, additional genes have been identified or hypothesised to be involved in the catabolism of sialic acid, predominately in *E. coli*. These supplementary genes are discussed below.

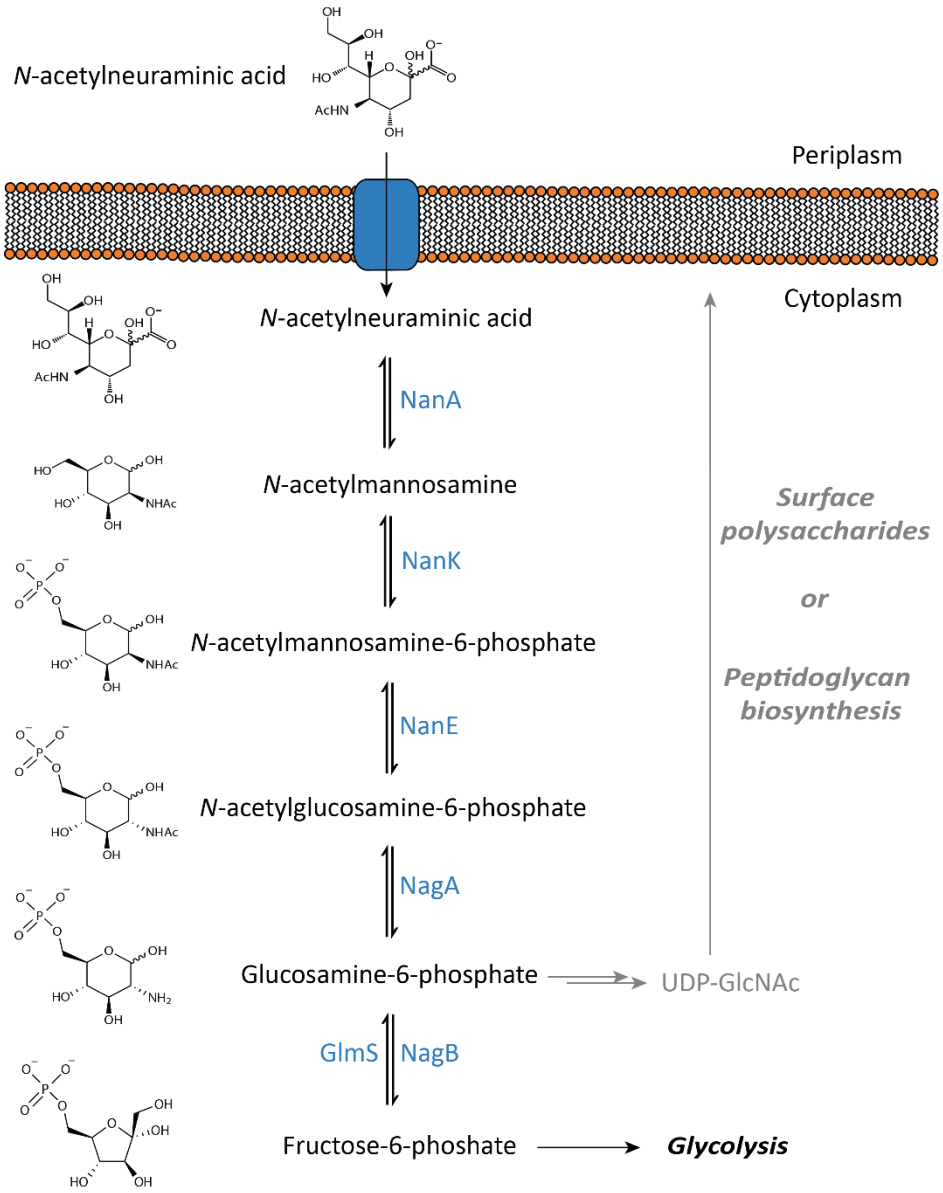


Figure 1.7 | Overview of sialic acid catabolism in bacteria. In brief, the degradation of sialic acid within bacteria is initiated by NanA, an enzyme that catalyses the retro-aldol cleavage of sialic acid to yield *N*-acetyl-D-mannosamine and pyruvate *via* a Schiff base intermediate. This retro-aldol reaction constitutes the first committed step of the degradation pathway. *N*-Acetyl-D-mannosamine is then phosphorylated at the C6 position by NanK, an ATP-dependent reaction to yield *N*-acetyl-D-mannosamine-6-phosphate and ADP. Thereafter, NanE inverts the stereochemistry of *N*-acetyl-D-mannosamine-6-phosphate in a reversible epimerisation reaction, generating *N*-acetyl-D-glucosamine-6-phosphate. Next, the acetyl group from *N*-acetyl-D-glucosamine-6-phosphate is subsequently removed by NagA, yielding glucosamine-6-phosphate. Finally, the enzyme NagB, catalyses the two-step isomerisation and deamination of glucosamine-6-phosphate to generate fructose-6-phosphate, which can then enter central metabolism *via* the glycolytic pathway. The reverse reaction is catalysed by glutamine-fructose-6-phosphate aminotransferase (GlmS). Here, glucosamine-6-phosphate can also yield cell wall precursors for peptidoglycan biosynthesis, or surface polysaccharides for poly-*N*-acetylglucosamine biosynthesis (shown in grey). The chemical structure of Neu5Ac and each catabolite along the sialic acid catabolic pathway is shown.

1.7.1 *N*-Acetylneuraminate lyase

Once sialic acid is imported into the cell through one of the dedicated transporters, *N*-acetylneuraminate lyase (NanA) is responsible for catalysing the primary and committed step in the catabolism of sialic acid (36). Through an aldol condensation via a Schiff base intermediate, *N*-acetylmannosamine and pyruvate is formed (93; 94). This aldolase activity is reversible (93) and was first identified over 45 years ago in a study by Nees and Schauer using *C. perfringens* (95).

NanA enzymes are well characterised across the bacterial pathogens, *C. perfringens* (95), *E. coli* (93; 94), *H. influenzae* (96), and *S. aureus* (97), which revealed a preference for the α-anomer of Neu5Ac (Figure 1.1A), over the more thermodynamically favourable β-anomer (22; 94). This stereoselective preference is a hallmark feature of lyase/ aldolase enzymes (98). Structurally, NanA enzymes display a characteristic $(\beta/\alpha)_8$-barrel domain architecture and present a tetrameric assembly (97; 99), which is essential for function (100). Owing to their stability and reversible chemistry, NanA enzymes have been successfully utilised in large-scale production of sialic acid for industry (101).

1.7.2 *N*-Acetylmannosamine kinase

N-Acetylmannosamine kinase (NanK) is the second enzyme in the core catabolic pathway for sialic acid. NanK catalyses the transfer of a phosphate moiety from ADP to the C6 position of *N*-acetylmannosamine to form *N*-acetylmannosamine-6-phosphate. NanK is a member of the repressor, open-reading frame, kinase (ROK) superfamily of proteins and contains an N-terminal adenosine-nucleotide binding domain, a zinc-binding domain, and a conserved arginine residue in the active site. This architecture is conserved and serves as a trademark for this family (102; 103). Overall, these two domains form a dimeric, butterfly-shaped architecture connected by a hinge region. This hinge is believed to facilitate catalysis by mediating the movement between an open and closed conformation upon substrate binding (104). Among ROK members, the zinc ion has been hypothesised to stabilise the active site (102). However, this binding site is absent in all reported NanK enzymes (105; 106), apart from *E. coli* which coordinates zinc predominantly *via* cysteine residues. This general absence has yet to be explained.

Structurally, the current understanding of NanK is limited to methicillin-resistant *S. aureus* (105) and *Fusobacterium nucleatum* (106), with only deposited coordinates of an *E. coli* and *Listeria monocytogenes* NanK (PDB ID 2AA4 and 4HTL, respectively). Consequently, our mechanistic understanding is based on the human bifunctional homolog UDP *N*-acetylglucosamine 2-epimerase/ *N*-acetylmannosamine kinase (102).

1.7.3 *N*-Acetylmannosamine-6-phosphate 2-epimerase

Following the phosphotransferase activity of NanK, the final enzyme of the Nan cluster, *N*-acetylmannosamine-6-phosphate 2-epimerase (NanE) catalyses the epimerisation of *N*-acetylmannosamine-6-phosphate to form *N*-acetylglucosamine-6-phosphate (107-109), inverting the stereochemistry at the C2 position.

Mechanistically, carbohydrate-based epimerases can facilitate this epimerisation through five different strategies: deprotonation/ re-protonation, carbon-carbon bond cleavage, nucleotide elimination, via formation of a transient keto intermediate and mutarotation (110; 111). Supported by proton-NMR kinetic experiments, the current understanding is that this mechanism employs the deprotonation/ re-protonation strategy, mediated via an enolate intermediate (109). However, for this strategy to be possible it relies on the substrate undergoing a 180° flip about a carbon-carbon bond. While this has been reported in other carbohydrate-based epimerases (112; 113), the molecular determinants in NanE has yet to be confirmed, and therefore the true epimerisation strategy remains unclear.

Structurally, NanE belongs to the triose-phosphate isomerase barrel superfamily, where by virtue the domain architecture of NanE contains a $(\beta/\alpha)_8$-barrel domain, which forms a dimeric assembly through an exchange of C-terminal α-helices (107-109). Together, this forms a helix-swapped interface. Furthermore, while NanE is conserved among bacteria, it shares no structural homology with human UDP-*N*-acetylglucosamine 2-epimerases. This observation implicates NanE as a viable target for the development of novel antimicrobial drugs.

1.7.4 *N*-Acetylglucosamine-6-phosphate deacetylase

N-Acetylglucosamine-6-phosphate deacetylase (NagA) catalyses the deacetylation of *N*-acetylglucosamine-6-phosphate to form glucosamine-6-phosphate, *via* process that utilises water and releases an acetate moiety (114). Although this reaction is the penultimate in the sialic acid catabolic pathway, it alternatively serves as the first committed step in the production of cell wall precursors for peptidoglycan biosynthesis, or surface polysaccharides for poly-*N*-acetylglucosamine biosynthesis (34; 35). This branch point (Figure 1.7) is essential for responding to the metabolic needs of the cell and to prevent the futile cycling of amino sugars (115). Importantly, this feature has warranted attention of NagA as a viable target for the developmemt of novel antimicrobial drugs (35; 116).

NagA belongs to the metal-dependent amidohydrolase superfamily and consists of two domains: a $(\beta/\alpha)_8$-barrel domain; and a small β-barrel domain (116). Enclosed by the $(\beta/\alpha)_8$-barrel domain, the

active site is reported to contain one or more divalent metal ions, including Fe^{2+}, Zn^{2+}, Cu^{2+} and Co^{2+} (117). This metal ion specificity varies between species due to different coordination motifs. The presence of these divalent metals ions is essential for activity, where they play a role in stabilising the acetate carbonyl oxygen, and the attacking nucleophile, while in the tetrahedral transition state (116). Among the NagA enzymes, the majority form a dimeric assembly (35; 118), although *E. coli* NagA is a tetramer in solution (117).

1.7.5 Glucosamine-6-phosphate deaminase

Following the deacetylation of *N*-acetyl-D-glucosamine-6-phosphate, glucosamine-6-phosphate deaminase (NagB) catalyses the final step in the sialic acid catabolic pathway, facilitating the isomerisation and deamination of glucosamine-6-phosphate to form fructose-6-phosphate and release ammonia. Importantly, fructose-6-phosphate can be directly fed into central metabolism as a glycolytic precursor (35; 119). Like NagA, NagB is also a branch point enzyme that determines the metabolic fate of glucosamine-6-phosphate. If the intracellular pool of amino sugars is low, glucosamine-6-phosphate can be formed from fructose-6-phosphate via the reverse reaction. This is catalysed by glutamine-fructose-6-phosphate aminotransferase, a process that utilises glutamine and releases glutamate (35; 120). However, high levels of ammonia have been reported to facilitate the reverse reaction, which is catalysed by NagB alone (121; 122).

NagB is classified in the aldose-ketose isomerase class of proteins, which characteristically catalyse the removal of a proton from the C2 position of the aldose and proton transfer to the C1 position of the ketose *via* a *cis*-enediol intermediate (123). Structurally, the domain architecture of NagB resembles a α/β-type Rossmann fold (124), while the quaternary assembly has been observed to vary between species. The majority of NagB enzymes display a monomeric assembly (119; 124), although *E. coli* NagB is reported as a hexamer (123), while *S. aureus* NagB is reported to be dimeric (35). Interestingly, only the *E. coli* hexamer exhibits allosteric regulation by its substrate glucosamine-6-phosphate (123), which raises questions about its physiological relevance between NagB homologs.

1.7.6 Supplementary enzymes of sialic acid catabolism

In addition to the *nan* and *nag* genes, several other genes have been identified as part of the Nan cluster in *E. coli*, which are hypothesised to be involved in the catabolism of sialic acid and provide support to the core catalytic pathway (22). Firstly, these include the genes encoded by the *nanCMS* operon, which include the outer membrane porin (NanC); the sialate mutarotase (NanM); and the 9-O-acetylated

Neu5Ac esterase (NanS). Introduced earlier, these genes function to uptake sialic acid across the outer membrane (48; 49), catalyse the periplasmic conversion of Neu5Ac to the thermodynamically favoured β-anomer (54), and permit growth 9-O-acetylated Neu5Ac (55; 56), respectively. Secondly, encoded by the *nanATEK-yhcH* operon, YhcH is hypothesised to function as an epimerase for *N*-glycolylneuraminic acid, on the basis of the crystal structure solved for this enzyme (125). Furthermore, encoded by the *yjhBC* operon, YjhB and YjhC are hypothesised to function as a permease and oxidoreductase, to uptake and process less common forms of sialic acid, respectively (22). However, no experimental evidence exists to support this hypothesis. To address this, the uncharacterised *yjhBC* operon will be investigated in Chapter Two and Chapter Three of this thesis.

1.7.7 The catabolism of sialic acid is a viable target for antimicrobial development

The human gastrointestinal tract is heavily populated by bacteria (126). As such, this environment is largely glucose-limiting and thus is unfavourable for invading pathogenic bacteria to colonise and establish infection within the human host (127). To overcome this bottleneck, pathogenic bacteria evolved the catabolic machinery necessary to utilise host-derived sialic acid as an alternate source of nutrition and gain a competitive advantage over neighbouring microbial residents (23; 128).

The ability to utilise host-derived sialic acid was first identified within *C. perfringens* (95) and has since been identified in *B. fragilis* (129), *E. coli* (24; 36), *H. influenzae* (39; 92), *S. aureus (130)*, *S. enterica* serovar Typhimurium (7), *V. cholerae* (23)*, V. vulnificus (127)*, and 27 other pathogenic bacteria, confined to the Gamma-Proteobacteria class and the Firmicutes phylum (7; 131). Knocking out this catabolic machinery *in vitro* resulted in the inability for these pathogens to grow on sialic acid as a sole source of nutrition, suggesting these enzymes serve the only pathway for sialic acid to be degraded (36; 127; 130). This was demonstrated *in vivo* using mouse models, where knocking out the *nanA* gene in *E. coli* and *V. cholerae,* significantly affected each respective pathogens ability to colonise and persist within the host (23; 24). Taken together, this highlights the importance of sialic acid utilisation in pathogenic bacteria, which is intimately linked to their fitness and survival within the human host.

Consequently, the sialic acid catabolic machinery is identified as a viable target for the development of novel antimicrobial therapeutics. Drug development could target the uptake of sialic acid across several sialic acid-specific transporters, which are unique to bacteria or one of the catabolic enzymes within the Nan cluster, such as the lyase (NanA) which catalyses the first committed step in the pathway. The dependency on the activity of this enzyme within pathogenic bacteria has been well documented across numerous *in vitro* and *in vivo* studies. With greater insight into the sialic acid-specific transporter proteins, this drug development to could be jointly targeted at both sialic acid transport and catabolism.

1.8 Regulation of sialic acid catabolism

Given the importance of sialic acid to the survival and persistence of pathogenic bacteria, it comes as no surprise that its catabolic pathway is regulated to ensure maximum efficiency and provide the pathogen a competitive advantage in glucose-limiting environments. The response to altered nutrient sources or an inopportune environment is most commonly controlled at the genetic level by transcriptional regulators. Here, expression of the catabolic machinery only occurs when the substrate is detected at levels that will ensure a sufficient metabolic return for it be utilised as a source of nutrition (Figure 1.8) (132; 133)—this is referred to as carbon catabolite repression (134). Without sufficient levels of the substrate, expression of the catabolic operon is switched off by the binding of a transcriptional regulator to the promoter region, which blocks the activity of RNA polymerase (135).

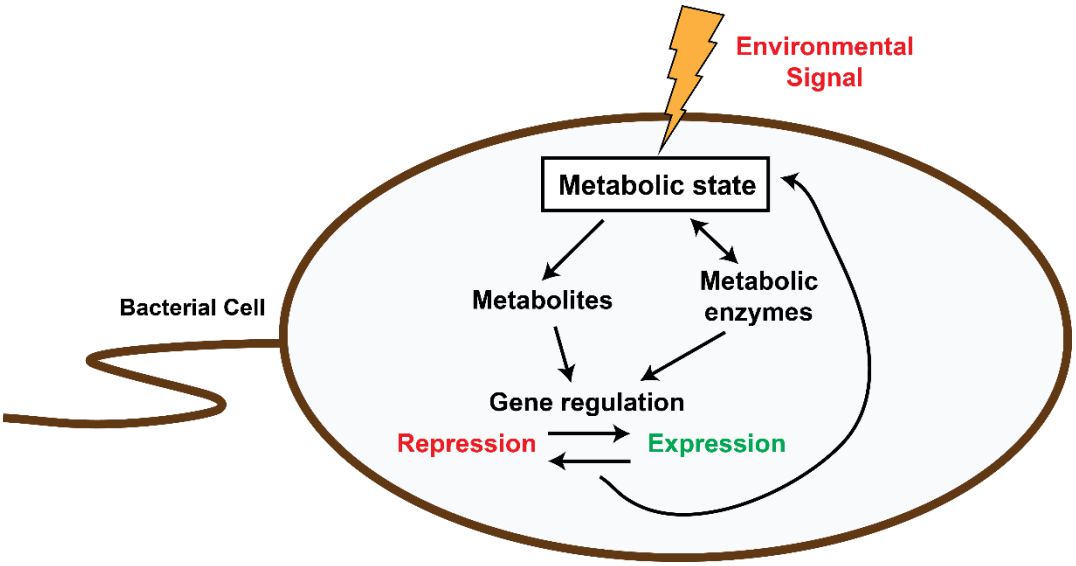


Figure 1.8 | Sensing nutrients in the environment. Bacteria selectively import and catabolise a wide variety of nutrients by processes that are tightly regulated in order to efficiently match the environment that the bacterium inhabits. However, the immediate environment can rapidly change and thus the ability to quickly adapt to these environmental signals (lightning bolt) provide bacteria with a competitive advantage. Mediated by transcriptional regulators, this process is achieved through changes in gene expression to upregulate the metabolic enzymes that are required to metabolise the available nutrient, such as sialic acid.

The gene expression of the sialic acid catabolic machinery is mediated by a transcriptional regulator that is referred to as the *N*-acetylneuraminic repressor, or herein abbreviated to NanR. In general, NanR functions as a gene repressor to negatively control the expression of the Nan cluster, which includes the sialic acid-specific transporter and the catabolic enzymes NanA, NanK and NanE (22). As NanR-mediated regulation is a feature of this thesis, a general introduction to the domain architecture, protein family and mechanism is presented below.

Conversely, the expression of the *nag* genes, which encode the Nag enzymes is controlled by the transcriptional regulator, NagC (115; 136). In addition, NagC activates the expression of the *glmUS* operon, which encodes the enzyme machinery to drive peptidoglycan biosynthesis (137). Together, regulation of these genes is essential for coordinating amino sugar metabolism in bacteria and to prevent futile cycling (115). While there is a limited molecular understanding of this NagC-mediated regulation, this was beyond the scope of this thesis.

Transcriptional regulation aside, catabolic pathways can also be regulated through a feedback mechanism, whereby catabolites within the pathway can influence flux through the pathway itself by controlling the activity of one or more enzymes by allostery (138). Interestingly, allostery does not appear to be a feature of the sialic acid catabolic pathway, as only the *E. coli* NagB hexamer has exhibited allosteric regulation by its substrate, glucosamine-6-phosphate (123).

1.8.1 The structural architecture of NanR transcriptional regulators

In general, NanR is comprised of two domains, which include an N-terminal DNA-binding domain and a C-terminal effector-binding domain (Figure 1.9) (22; 139). The N-terminal domain binds DNA with high affinity to the promoter region of the catabolic operon, through a helix-turn-helix motif This is a common DNA-binding motif amongst bacterial transcriptional regulators, which is defined by a tight tri-helical structure, where the third α-helix is often referred to as the 'recognition helix' (Figure 1.10A), as it interacts directly with the major groove of the DNA through electrostatic interaction (140; 141).

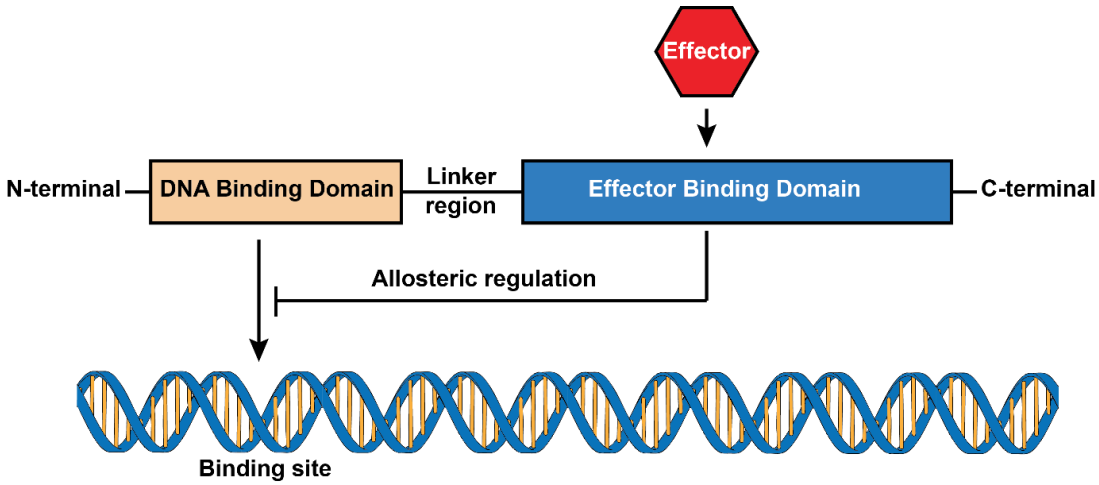


Figure 1.9 | Cartoon overview of the sialoregulator NanR. The N-terminal DNA-binding domain (beige) binds DNA with high affinity to a specific binding site on the catabolic operon. The binding of an effector molecule (red) to the C-terminal domain (blue) propagates a conformational change *via* a linker region to facilitate a change in activity of the repressor, through allosteric regulation.

By contrast, the C-terminal domain forms the oligomerisation interface and propagates conformational changes upon binding of an effector that decrease the affinity for DNA (derepression), or in some cases increase the affinity for DNA (activation) (22; 142). Allosteric regulation is a trademark of bacterial transcriptional regulators allowing them to serve as molecular switches, switching gene expression on or off (143). The factors that facilitate allosteric regulation in NanR are discussed in more detail below.

1.8.2 The protein family classification of NanR transcriptional regulators

Encoded by the *nanR* gene, NanR has been identified across a range of bacteria, including *E. coli* (144), *H. influenzae* (145), *S. aureus* (130), *S. pneumoniae* (146) and *V. vulnificus* (139), where it functions as repressor to negatively control the gene expression of sialic acid catabolic machinery in the absence of this alternative nutrient.

Surprisingly, given this overall function, NanR is annotated in two separate protein families, including the RpiR family of transcriptional regulators and the GntR family of transcriptional regulators (22; 144). While most pathogens contain a RpiR-type NanR, Enterobacteriaceae encode a NanR that belongs to the GntR superfamily, one of the most widespread families of transcriptional regulators that regulate gene expression through allostery upon binding to metabolites (142).

There are two main differences between the RpiR and GntR family of transcriptional regulators. Firstly, in comparison with RpiR family members, the helix-turn-helix motif of GntR family members is characterised by an additional 'wing' that interacts with the minor groove of DNA to provide additional specificity (Figure 1.10B). This 'wing' is formed by a small loop, with two short β-strands following the canonical tri-helical structure (147). Secondly, in addition to regulating the Nan cluster, the GntR-type NanR has been shown to control the expression of two other operons: 1) *nanCMS*, responsible for outer-membrane transport and periplasmic processing; and 2) *yjhBC*, whose function is currently undefined, but is investigated in Chapter Two. This coordinated regulation of multiple operons is known as the sialoregulon and is only present in *E. coli* (144). This unique regulation of sialic acid catabolism will be investigated in Chapter Five and Six.

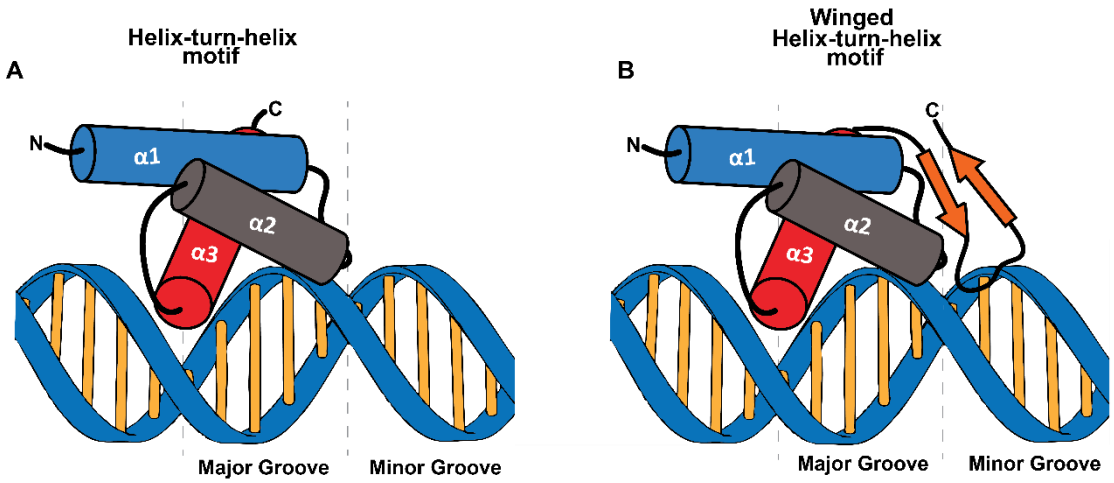


Figure 1.10 | Cartoon overview of the helix-turn-helix and the winged-helix-turn-helix motif. A) The canonical helix-turn-helix motif. Structurally, this motif is defined by a tight tri-helical structure, where the third α-helix (α3 in red) is often referred to as the 'recognition helix', as it interacts directly with the major groove of the DNA through electrostatic interaction. **B)** The winged-helix-turn-helix variant of the canonical DNA-binding motif. Structurally, this is characterised by an additional 'wing' that interacts with the minor groove of DNA to provide additional specificity. For clarity, both major and minor DNA grooves are highlighted.

1.8.3 Allosteric mechanism of NanR transcriptional regulators

Allostery plays a critical role in the function of NanR transcriptional regulators (142). This process is driven by the binding of an effector molecule to the C-terminal effector-binding domain, which propagates a conformation change through the protein to elicit a change in DNA-binding affinity. The effector molecule that mediates the activity of NanR is most commonly a degradation product of the sialic acid catabolic pathway (148), such as *N*-acetyl-D-mannosamine-6-phosphate, which has been identified as the inducer molecule for NanR from *V. vulnificus* (149) and *S. aureus* (130). Although, as an exception, sialic acid has been implicated as the inducer molecule in *E. coli* (144; 148).

The ability to attenuate the expression of this catabolic machinery on and off is essential for the bacterial pathogen to gain a competitive advantage in the human host. Furthermore, constitutive expression of the sialic acid-specific transporters would be metabolically unfavourable for the bacterium, resulting in undue stress on an already crowded membrane environment. Thus, gene expression is only induced when the signal (effector) is present. Some authors believe cytoplasmic accumulation of sialic acid to high levels is toxic, inhibiting growth (36), while others provide contradictory evidence (127). Hence, further investigation is required to understand this notion. Toxicity aside, the ability to respond rapidly to an environmental change is essential in order to utilise sialic acid as a source of nutrition and avoid futile cycling.

To better understand the allosteric mechanism, the availability of high-resolution structures that capture the repressor in different complex states (effector-free, effector-bound, and DNA-bound) is essential to map any conformational changes that occur (150). Traditionally, X-ray crystallography has been an invaluable tool to characterise these conformations that allow bacterial transcriptional regulators to serve as molecular switches (143). In the GntR superfamily, the common feature that alleviates the interaction with DNA is a large displacement of the N-terminal DNA-binding domains when an effector molecule binds (142; 151; 152). This mechanism is illustrated in Figure 1.11 to provide a model for allosteric regulation in NanR.

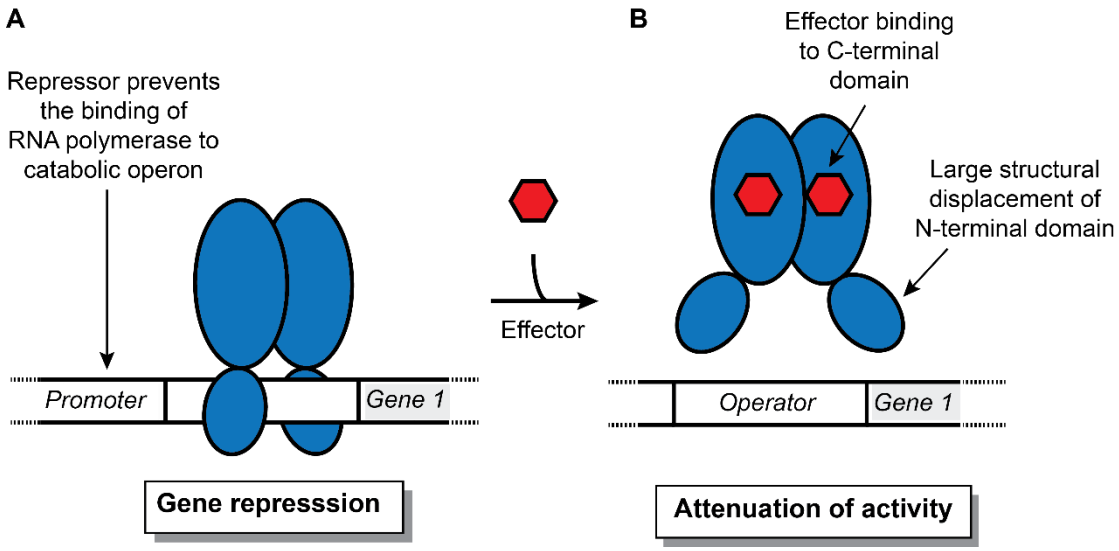


Figure 1.11 | Cartoon overview for allosteric regulation in the GntR superfamily. A) To regulate the gene expression of catabolic enzymes, GntR family members interact with the operator site of the respective catabolic operon. When bound to DNA, the repressor prevents the binding of RNA polymerase. **B)** Upon effector binding (red hexagon) to the C-terminal domain (large blue oval), a large structural displacement in the N-terminal DNA binding domain (small blue oval), leads to an attenuation of DNA-binding activity.

Among NanR transcriptional regulators, allosteric regulation is poorly understood as there is only a single crystal structure solved (Figure 1.12)—the 1.9 Å structure of *V. vulnificus* in complex with its allosteric effector, *N*-acetyl-D-mannosamine-6-phosphate (PDB ID - 4IVN). This RpiR-type NanR is believed to form a dimeric assembly, where DNA is hypothesised to bind between the monomers in an arched tunnel-like cavity, mediated by positively charged residues (139). This data allowed the authors to propose the mechanism for the transcriptional regulation of sialic acid catabolism. However, interpreted from negative stain transmission electron microscopy data and crystal symmetry mates, the dimer interface of their functional regulator is unconvincing, as is formed by only four hydrogen bonds

and buries less than 3% of the accessible surface area (Figure 1.12). Together, these features would support a very weak and transient oligomeric assembly. Moreover, these authors had previously demonstrated that *V. vulnificus* NanR adopts a hexameric assembly in solution (149), as opposed to a dimer. Therefore, the active DNA bound conformation and true oligomeric assembly of RpiR-type NanR has yet to determined. This open question will be addressed in Chapter Four for the RpiR-type NanR from *S. pneumoniae*.

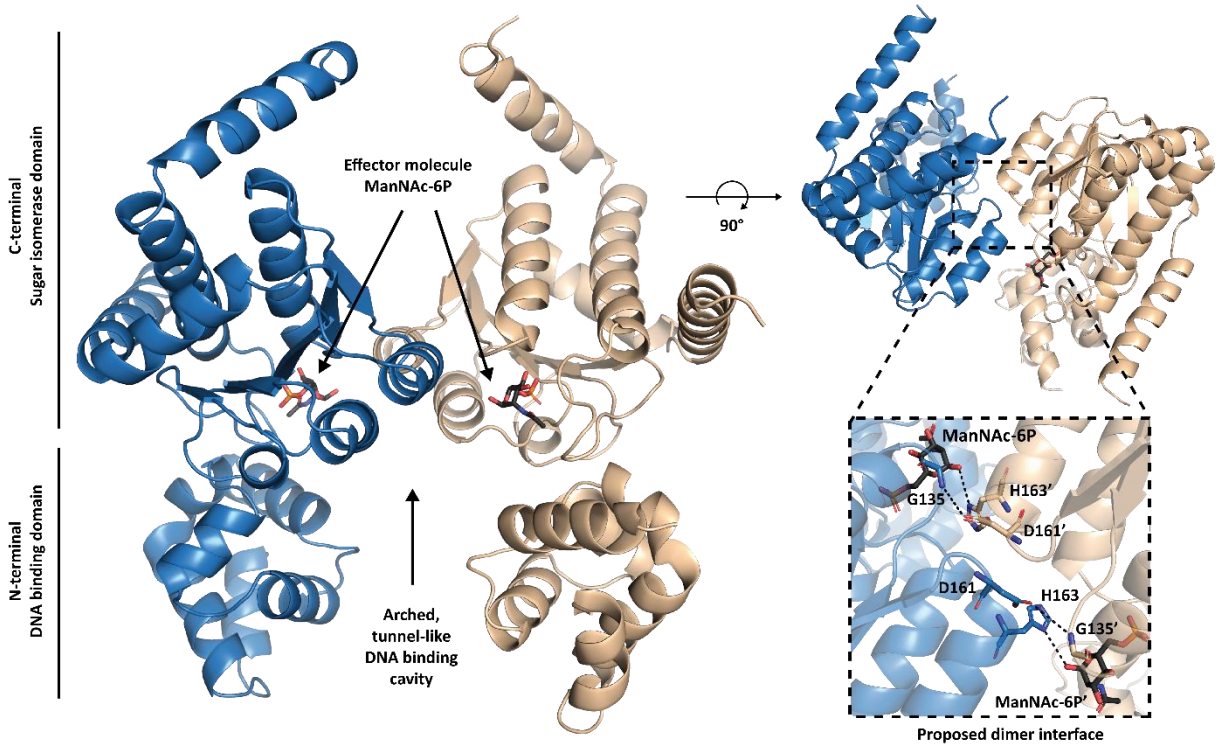


Figure 1.12 | Crystal structure of the *V. vulnificus* NanR. The left panel presents the proposed dimeric assembly for *V. vulnificus* RpiR-type NanR in its 'active state', in complex with the *N*-acetyl-D-mannosamine-6-phopshate (ManNAc-6P) effector molecule (sticks). Highlighted is the arched, tunnel-like DNA-binding cavity. The right panel presents a top view of the transcriptional regulator, where the insert highlights the four hydrogen bonds that form the dimeric interface—these are formed between residues, D161-G135 and H163-ManNAc-6P. This figure was generated using PyMOL and the PDB entry 4IVN (139).

1.9 Research Aims

The overall aim of this research is to develop a molecular understanding of sialic acid gene regulation in bacteria and investigate the biological role of the uncharacterised *yjhBC* operon in sialic acid catabolism. Despite a number of groups addressing bacterial sialoregulation around the world, surprisingly little structural and biophysical data exists to elucidate how the transcriptional regulator functions.

Chapter Two tests the hypothesis that the functional role of the *E. coli yjhBC* operon is to import and process the less common derivatives of sialic acid. This work explores the biological role of *E. coli* YjhC placed in the Gfo/Idh/MocA family of oxidoreductases. Here, a biophysical characterisation and structural analysis are presented, highlighting several sialic acid and closely related variants are potential substrates for YjhC.

Chapter Three expands the investigation of the *E. coli yjhBC* operon and aims to understand the function of *E. coli* YjhB. This work describes the overexpression trials of this membrane protein and an *in silico* characterisation to deduce regions of functional importance.

Chapter Four provides a comprehensive study of *S. pneumoniae NanR*, a member of the RpiR family of transcriptional regulators. This chapter describes the purification, reports the binding kinetics, and presents novel structural data in solution. Together, this is a substantial step towards structurally and functionally understanding how these RpiR-like family of regulators control gene expression.

Chapter Five thoroughly investigates the GntR-type NanR from *E. coli*, reports the first crystal structure of a GntR-type sialoregulator and presents a model for the DNA-bound conformation to near atomic resolution using cryo-electron microscopy. Through collaboration with Professor Borries Demeler (University of Lethbridge), this thesis confirms the stoichiometry of DNA-binding using novel multi-wavelength sedimentation velocity experiments, which I designed, conducted and analysed. Excitingly, this work provides a detailed molecular understanding of how NanR regulates gene expression and reveals an unusual mechanism of cooperativity that drives protein-DNA hetero-complex formation.

Chapter Six provides supplementary information to support the model for gene regulation in *E. coli*. This work reports the thermal stability of NanR, describes the attempts at crystallisation of this transcriptional regulator with DNA and provides a background for the novel multi-wavelength sedimentation velocity experiments used to define to the stoichiometry of the system.

To summarise, these data enhance our understanding of bacterial sialoregulation, contributes to a better grasp of pathogenicity and informs the design and development of novel antimicrobial therapeutics.

1.10 Chapter References

1. Schauer, R. (2000). Achievements and challenges of sialic acid research. *Glycoconjugate Journal, 17*, 485-499.
2. Varki, A. (2008). Sialic acids in human health and disease. *Trends in Molecular Medicine, 14*, 351-360.
3. Angata, T., & Varki, A. (2002). Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. *Chemical Reviews, 102*, 439-469.
4. Varki, A. (1992). Diversity in the sialic acids. *Glycobiology, 2*, 25-40.
5. Varki, A. (1997). Sialic acids as ligands in recognition phenomena. *Federation of American Societies for Experimental Biology Journal, 11*, 248-255.
6. Vimr, E. R., Kalivoda, K. A., Deszo, E. L., & Steenbergen, S. M. (2004). Diversity of microbial sialic acid metabolism. *Microbiology and Molecular Biology Reviews, 68*, 132-153.
7. Almagro-Moreno, S., & Boyd, E. F. (2009). Insights into the evolution of sialic acid catabolism among bacteria. *BioMed Central Evolutionary Biology, 9*, 118.
8. Schauer, R. (2004). Sialic acids: fascinating sugars in higher animals and man. *Zoology, 107*, 49-64.
9. Dwek, R. A. (1996). Glycobiology: Toward understanding the function of sugars. *Chemical Reviews, 96*, 683-720.
10. Kleene, R., & Schachner, M. (2004). Glycans and neural cell interactions. *Nature Reviews Neuroscience, 5*, 195-208.
11. Varki, A. (1993). Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology, 3*, 97-130.
12. Traving, C., & Schauer, R. (1998). Structure, function and metabolism of sialic acids. *Cellular and Molecular Life Sciences, 54*, 1330-1349.
13. Kelm, S., & Schauer, R. (1997). Sialic acids in molecular and cellular interactions (Vol. 175, pp. 137-240). San Diego: Elsevier Academic Press Inc.
14. Varki, A. (2007). Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. *Nature, 446*, 1023-1029.
15. Blix, F. G., Gottschalk, A., & Klenk, E. (1957). Proposed Nomenclature in the Field of Neuraminic and Sialic Acids. *Nature, 179*, 1088-1088.
16. Baos, S. C., Phillips, D. B., Wildling, L., McMaster, T. J., & Berry, M. (2012). Distribution of sialic acids on mucins and gels: A defense mechanism. *Biophysical Journal, 102*, 176-184.
17. Robbe, C., Capon, C., Coddeville, B., & Michalski, J.-C. (2004). Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract. *The Biochemical Journal, 384*, 307-316.
18. Lewis, A. L., & Lewis, W. G. (2012). Host sialoglycans and bacterial sialidases: a mucosal perspective. *Cellular Microbiology, 14*, 1174-1182.
19. Ho, J. J., Cheng, S., & Kim, Y. S. (1995). Access to peptide regions of a surface mucin (MUC1) is reduced by sialic acids. *Biochemical and Biophysical Research Communications, 210*, 866-873.
20. Dethlefsen, L., McFall-Ngai, M., & Relman, D. A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature, 449*, 811-818.
21. Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell, 124*, 837-848.
22. Vimr, E. R. (2013). Unified theory of bacterial sialometabolism: how and why bacteria metabolise host sialic acids. *International Scholarly Research Notices Microbiology, 2013*, 816713.
23. Almagro-Moreno, S., & Boyd, E. F. (2009). Sialic acid catabolism confers a competitive advantage to pathogenic *Vibrio cholerae* in the mouse intestine. *Infection and Immunity, 77*, 3807-3816.
24. Chang, D. E., Smalley, D. J., Tucker, D. L., Leatham, M. P., Norris, W. E., Stevenson, S. J., Anderson, A. B., Grissom, J. E., Laux, D. C., Cohen, P. S., & Conway, T. (2004). Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 7427-7432.
25. Corfield, A. P., Veh, R. W., Wember, M., Michalski, J. C., & Schauer, R. (1981). The release of *N*-acetyl- and *N*-glycolloyl-neuraminic acid from soluble complex carbohydrates and erythrocytes by bacterial, viral and mammalian sialidases. *Biochemical Journal, 197*, 293-299.
26. Bouchet, V., Hood, D. W., Li, J., Brisson, J. R., Randle, G. A., Martin, A., Li, Z., Goldstein, R., Schweda, E. K., Pelton, S. I., Richards, J. C., & Moxon, E. R. (2003). Host-derived sialic acid is incorporated into *Haemophilus influenzae* lipopolysaccharide and is a major virulence factor in experimental otitis media. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 8898-8903.

27. Sohanpal, B. K., El-Labany, S., Lahooti, M., Plumbridge, J. A., & Blomfield, I. C. (2004). Integrated regulatory responses of fimB to *N*-acetylneuraminic (sialic) acid and GlcNAc in *Escherichia coli* K-12. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 16322-16327.
28. Shakhnovich, E. A., King, S. J., & Weiser, J. N. (2002). Neuraminidase expressed by *Streptococcus pneumoniae* desialylates the lipopolysaccharide of N*eisseria meningitidis* and *Haemophilus influenzae*: a paradigm for interbacterial competition among pathogens of the human respiratory tract. *Infection and Immunity, 70*, 7161-7164.
29. Taeko, M., & Kazunori, Y. (2012). Mammalian sialidases: Physiological and pathological roles in cellular functions. *Glycobiology, 22*, 880-896.
30. Burnaugh, A. M., Frantz, L. J., & King, S. J. (2008). Growth of *Streptococcus pneumoniae* on Human Glycoconjugates Is Dependent upon the Sequential Activity of Bacterial Exoglycosidases. *Journal of Bacteriology, 190*, 221-230.
31. Mizan, S., Henk, A., Stallings, A., Maier, M., & Lee, M. D. (2000). Cloning and Characterization of Sialidases with 2-6′ and 2-3′ Sialyl Lactose Specificity from *Pasteurella multocida*. *Journal of Bacteriology, 182*, 6874-6883.
32. Sillanaukee, P., Ponnio, M., & Jaaskelainen, I. P. (1999). Occurrence of sialic acids in healthy humans and different disorders. *European Journal of Clinical Investigation, 29*, 413-425.
33. McDonald, N. D., Lubin, J. B., Chowdhury, N., & Boyd, E. F. (2016). Host-Derived Sialic Acids Are an Important Nutrient Source Required for Optimal Bacterial Fitness in vivo. *MBio, 7*, e02237-02215.
34. Severi, E., Hood, D. W., & Thomas, G. H. (2007). Sialic acid utilisation by bacterial pathogens. *Microbiology, 153*, 2817-2822.
35. Davies, J. S., Coombes, D., Horne, C. R., Pearce, F. G., Friemann, R., North, R. A., & Dobson, R. C. J. (2019). Functional and solution structure studies of amino sugar deacetylase and deaminase enzymes from *Staphylococcus aureus*. *FEBS Letters, 593*, 52-66.
36. Vimr, E. R., & Troy, F. A. (1985). Identification of an inducible catabolic system for sialic acids (nan) in *Escherichia coli*. *Journal of Bacteriology, 164*, 845-853.
37. Rohmer, L., Hocquet, D., & Miller, S. I. (2011). Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends in Microbiology, 19*, 341-348.
38. Schmidt, H., & Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clinical and Microbiology Reviews, 17*, 14-56.
39. Severi, E., Randle, G., Kivlin, P., Whitfield, K., Young, R., Moxon, R., Kelly, D., Hood, D., & Thomas, G. H. (2005). Sialic acid transport in *Haemophilus influenzae* is essential for lipopolysaccharide sialylation and serum resistance and is dependent on a novel tripartite ATP-independent periplasmic transporter. *Molecular Microbiology, 58*, 1173-1185.
40. Severi, E., Hosie, A. H. F., Hawkhead, J. A., & Thomas, G. H. (2010). Characterisation of a novel sialic acid transporter of the sodium solute symporter (SSS) family and *in vivo* comparison with known bacterial sialic acid transporters. *Federation of European Microbiological Societies Microbiology Letters, 304*, 47-54.
41. Martinez, J., Steenbergen, S., & Vimr, E. (1995). Derived structure of the putative sialic acid transporter from *Escherichia coli* predicts a novel sugar permease domain. *Journal of Bacteriology, 177*, 6005-6010.
42. Wahlgren, W. Y., Dunevall, E., North, R. A., Paz, A., Scalise, M., Bisignano, P., Bengtsson-Palme, J., Goyal, P., Claesson, E., Caing-Carlsson, R., Andersson, R., Beis, K., Nilsson, U. J., Farewell, A., Pochini, L., Indiveri, C., Grabe, M., Dobson, R. C. J., Abramson, J., Ramaswamy, S., & Friemann, R. (2018). Substrate-bound outward-open structure of a Na(+)-coupled sialic acid symporter reveals a new Na(+) site. *Nature Communications, 9*, 1753.
43. Post, D. M., Mungur, R., Gibson, B. W., & Munson, R. S., Jr. (2005). Identification of a novel sialic acid transporter in *Haemophilus ducreyi*. *Infection and Immunity, 73*, 6727-6735.
44. Chowdhury, N., Norris, J., McAlister, E., Lau, S. Y., Thomas, G. H., & Boyd, E. F. (2012). The VC1777-VC1779 proteins are members of a sialic acid-specific subfamily of TRAP transporters (SiaPQM) and constitute the sole route of sialic acid uptake in the human pathogen *Vibrio cholerae*. *Microbiology, 158*, 2158-2167.
45. Allen, S., Zaleski, A., Johnston, J. W., Gibson, B. W., & Apicella, M. A. (2005). Novel sialic acid transporter of *Haemophilus influenzae*. *Infection and Immunity, 73*, 5291-5300.
46. Mulligan, C., Leech, A. P., Kelly, D. J., & Thomas, G. H. (2012). The membrane proteins SiaQ and SiaM form an essential stoichiometric complex in the sialic acid tripartite ATP-independent periplasmic (TRAP) transporter SiaPQM (VC1777–1779) from *Vibrio cholerae*. *Journal of Biological Chemistry, 287*, 3598-3608.

47. North, R. A., Horne, C. R., Davies, J. S., Remus, D. M., Muscroft-Taylor, A. C., Goyal, P., Wahlgren, W. Y., Ramaswamy, S., Friemann, R., & Dobson, R. C. J. (2018). "Just a spoonful of sugar...": import of sialic acid across bacterial cell membranes. *Biophysical Reviews, 10*, 219-227.
48. Wirth, C., Condemine, G., Boiteux, C., Berneche, S., Schirmer, T., & Peneff, C. M. (2009). NanC crystal structure, a model for outer-membrane channels of the acidic sugar-specific KdgM porin family. *Journal of Molecular Biology, 394*, 718-731.
49. Condemine, G., Berrier, C., Plumbridge, J., & Ghazi, A. (2005). Function and expression of an *N*-acetylneuraminic acid-inducible outer membrane channel in *Escherichia coli*. *Journal of Bacteriology, 187*, 1959-1965.
50. Roy, S., Douglas, C. W., & Stafford, G. P. (2010). A novel sialic acid utilization and uptake system in the periodontal pathogen *Tannerella forsythia*. *Journal of Bacteriology, 192*, 2285-2293.
51. Phansopa, C., Roy, S., Rafferty, J. B., Douglas, C. W., Pandhal, J., Wright, P. C., Kelly, D. J., & Stafford, G. P. (2014). Structural and functional characterization of NanU, a novel high-affinity sialic acid-inducible binding protein of oral and gut-dwelling Bacteroidetes species. *Biochemistry, 458*, 499-511.
52. Galdiero, S., Falanga, A., Cantisani, M., Tarallo, R., della Pepa, M. E., D'Oriano, V., & Galdiero, M. (2012). Microbe-Host interactions: Structure and role of Gram-negative bacterial Porins. *Current Protein and Peptide Science, 13*, 843-854.
53. Beveridge, T. J. (1999). Structures of gram-negative cell walls and their derived membrane vesicles. *Journal of Bacteriology, 181*, 4725-4733.
54. Severi, E., Müller, A., Potts, J. R., Leech, A., Williamson, D., Wilson, K. S., & Thomas, G. H. (2008). Sialic acid mutarotation is catalyzed by the *Escherichia coli* β-propeller protein YjhT. *Journal of Biological Chemistry, 283*, 4841-4849.
55. Steenbergen, S. M., Jirik, J. L., & Vimr, E. R. (2009). YjhS (NanS) Is Required for *Escherichia coli* To Grow on 9-O-Acetylated *N*-Acetylneuraminic Acid. *Journal of Bacteriology, 191*, 7134-7139.
56. Rangarajan, E. S., Ruane, K. M., Proteau, A., Schrag, J. D., Valladares, R., Gonzalez, C. F., Gilbert, M., Yakunin, A. F., & Cygler, M. (2011). Structural and enzymatic characterization of NanS (YjhS), a 9-O-Acetyl N-acetylneuraminic acid esterase from *Escherichia coli* O157:H7. *Protein Science, 20*, 1208-1219.
57. Johnston, J. W., Shamsulddin, H., Miller, A.-F., & Apicella, M. A. (2010). Sialic acid transport and catabolism are cooperatively regulated by SiaR and CRP in nontypeable *Haemophilus influenzae*. *BMC Microbiology, 10*, 240-240.
58. Marion, C., Aten, A. E., Woodiga, S. A., & King, S. J. (2011). Identification of an ATPase, MsmK, which energises multiple carbohydrate ABC transporters in *Streptococcus pneumoniae*. *Infection and Immunity, 79*, 4193-4200.
59. Quistgaard, E. M., Low, C., Guettou, F., & Nordlund, P. (2016). Understanding transport by the major facilitator superfamily (MFS): structures pave the way. *Nature Reviews Molecular Cell Biology, 17*, 123-132.
60. Sun, L., Zeng, X., Yan, C., Sun, X., Gong, X., Rao, Y., & Yan, N. (2012). Crystal structure of a bacterial homologue of glucose transporters GLUT1-4. *Nature, 490*, 361-366.
61. Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R., & Iwata, S. (2003). Structure and mechanism of the lactose permease of *Escherichia coli*. *Science, 301*, 610-615.
62. Huang, Y., Lemieux, M. J., Song, J., Auer, M., & Wang, D. N. (2003). Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science, 301*, 616-620.
63. Wright, E. M., Loo, D. D. F., Hirayama, B. A., & Turk, E. (2004). Surprising versatility of Na+-glucose cotransporters: SLC5. *Physiology, 19*, 370-376.
64. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y., & Gouaux, E. (2005). Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters. *Nature, 437*, 215-223.
65. Higgins, C. F. (1992). ABC Transporters - from microorganisms to man. *Annual Review of Cell Biology, 8*, 67-113.
66. Jones, P. M., & George, A. M. (2004). The ABC transporter structure and mechanism: perspectives on recent research. *Cellular and Molecular Life Science, 61*, 682-699.
67. Saurin, W., Hofnung, M., & Dassa, E. (1999). Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *Journal of Molecular Evolution, 48*, 22-41.
68. Biemans-Oldehinkel, E., Doeven, M. K., & Poolman, B. (2006). ABC transporter architecture and regulatory roles of accessory domains. *FEBS Letters, 580*, 1023-1035.
69. Rees, D. C., Lewinson, O., & Johnson, E. (2009). ABC transporters: the power to change. *Nature Reviews Molecular Cell Biology, 10*, 218-227.

70. Davidson, A. L., & Maloney, P. C. (2007). ABC transporters: how small machines do a big job. *Trends in Microbiology, 15*, 448-455.
71. Kelly, D. J., & Thomas, G. H. (2001). The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *Federation of European Microbiological Societies Microbiology Reviews, 25*, 405-424.
72. Gangi Setty, T., Mowers, J. C., Hobbs, A. G., Maiya, S. P., Syed, S., Munson, R. S., Jr., Apicella, M. A., & Subramanian, R. (2018). Molecular characterization of the interaction of sialic acid with the periplasmic binding protein from *Haemophilus ducreyi*. *Journal of Biological Chemistry, 293*, 20073-20084.
73. Marion, C., Burnaugh, A. M., Woodiga, S. A., & King, S. J. (2011). Sialic acid transport contributes to pneumococcal colonisation. *Infection and Immunity, 79*, 1262-1269.
74. Bidossi, A., Mulas, L., Decorosi, F., Colomba, L., Ricci, S., Pozzi, G., Deutscher, J., Viti, C., & Oggioni, M. R. (2012). A functional genomics approach to establish the complement of carbohydrate transporters in *Streptococcus pneumoniae*. *Public Library of Science One, 7*, e33320.
75. Mulligan, C., Fischer, M., & Thomas, G. H. (2011). Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea. *Federation of European Microbiological Societies Microbiology Reviews, 35*, 68-86.
76. Rosa, L. T., Bianconi, M. E., Thomas, G. H., & Kelly, D. J. (2018). Tripartite ATP-Independent Periplasmic (TRAP) Transporters and Tripartite Tricarboxylate Transporters (TTT): From Uptake to Pathogenicity. *Frontiers in Cellular and Infection Microbiology, 8*, 33.
77. Steenbergen, S. M., Lichtensteiger, C. A., Caughlan, R., Garfinkle, J., Fuller, T. E., & Vimr, E. R. (2005). Sialic acid metabolism and systemic pasteurellosis. *Infection and Immunity, 73*, 1284-1294.
78. Tatum, F. M., Tabatabai, L. B., & Briggs, R. E. (2009). Sialic acid uptake is necessary for virulence of *Pasteurella multocida* in turkeys. *Microbial Pathogenesis, 46*, 337-344.
79. Lubin, J. B., Kingston, J. J., Chowdhury, N., & Boyd, E. F. (2012). Sialic acid catabolism and transport gene clusters are lineage specific in *Vibrio vulnificus*. *Applied and Environmental Microbiology, 78*, 3407-3415.
80. Johnston, J. W., Coussens, N. P., Allen, S., Jon, C. D. H., Turner, K. H., Zaleski, A., Ramaswamy, S., Gibson, B. W., & Apicella, M. A. (2008). Characterization of the *N*-Acetyl-5-neuraminic Acid-binding Site of the Extracytoplasmic Solute Receptor (SiaP) of Nontypeable *Haemophilus influenzae* Strain 2019. *Journal of Biological Chemistry, 283*, 855-865.
81. Müller, A., Severi, E., Mulligan, C., Watts, A. G., Kelly, D. J., Wilson, K. S., Wilkinson, A. J., & Thomas, G. H. (2006). Conservation of structure and mechanism in primary and secondary transporters exemplified by SiaP, a sialic acid binding virulence factor from *Haemophilus influenzae*. *Journal of Biological Chemistry, 281*, 22212-22222.
82. Gangi Setty, T., Cho, C., Govindappa, S., Apicella, M. A., & Ramaswamy, S. (2014). Bacterial periplasmic sialic acid-binding proteins exhibit a conserved binding site. *Acta Crystallographica Section D, 70*, 1801-1811.
83. Oh, B. H., Pandit, J., Kang, C. H., Nikaido, K., Gokcen, S., Ames, G. F., & Kim, S. H. (1993). Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand. *Journal of Biological Chemistry, 268*, 11348-11355.
84. Felder, C. B., Graul, R. C., Lee, A. Y., Merkle, H. P., & Sadee, W. (1999). The Venus flytrap of periplasmic binding proteins: an ancient protein module present in multiple drug receptors. *AAPS PharmSciTech, 1*, E2.
85. Schauer, R. (2009). Sialic acids as regulators of molecular and cellular interactions. *Current opinion in structural biology, 19*, 507-514.
86. Vimr, E., & Lichtensteiger, C. (2002). To sialylate, or not to sialylate: that is the question. *Trends in Microbiology, 10*, 254-257.
87. Haines-Menges, B. L., Whitaker, W. B., Lubin, J. B., & Boyd, E. F. (2015). Host Sialic Acids: A Delicacy for the Pathogen with Discerning Taste. *Microbiol Spectr, 3*.
88. Chou, W. K., Hinderlich, S., Reutter, W., & Tanner, M. E. (2003). Sialic acid biosynthesis: Stereochemistry and mechanism of the reaction catalyzed by the mammalian UDP-N-acetylglucosamine 2-epimerase. *Journal of the American Chemical Society, 125*, 2455-2461.
89. Vann, W. F., Tavarez, J. J., Crowley, J., Vimr, E., & Silver, R. P. (1997). Purification and characterisation of the *Escherichia coli* K1 *neu*B gene product *N*-acetylneuraminic acid synthetase. *Glycobiology, 7*, 697-701.
90. Shell, D. M., Chiles, L., Judd, R. C., Seal, S., & Rest, R. F. (2002). The neisseria lipooligosaccharide-specific alpha-2,3-sialyltransferase is a surface-exposed outer membrane protein. *Infection and Immunity, 70*, 3744-3751.
91. Previato, J. O., Andrade, A. F., Pessolani, M. C., & Mendonca-Previato, L. (1985). Incorporation of sialic acid into *Trypanosoma cruzi* macromolecules. A proposal for a new metabolic route. *Molecular Biochemistry and Parasitology, 16*, 85-96.

92. Vimr, E., Lichtensteiger, C., & Steenbergen, S. (2000). Sialic acid metabolism's dual function in Haemophilus influenzae. *Molecular Microbiology, 36*, 1113-1123.
93. Izard, T., Lawrence, M. C., Malby, R. L., Lilley, G. G., & Colman, P. M. (1994). The three-dimensional structure of *N*-acetylneuraminate lyase from *Escherichia coli*. *Structure, 2*, 361-369.
94. Lawrence, M., Barbosa, J., Smith, B., Hall, N., Pilling, P., Ooi, H., & Marcuccio, S. (1997). Structure and mechanism of a sub-family of enzymes related to *N*-acetylneuraminate lyase. *Journal of Molecular Biology, 266*, 381-399.
95. Nees, S., & Schauer, R. (1974). Induction of neuraminidase from *Clostridium perfringens* and the cooperation of this enzyme with acylneuraminate pyruvate lyase. *Behring Institute Mitteilungen, 55*, 68-78.
96. Barbosa, J., Smith, B., DeGori, R., Ooi, H., Marcuccio, S., Campi, E., Jackson, W., Brossmer, R., Sommer, M., & Lawrence, M. (2000). Active site modulation in the *N*-acetylneuraminate lyase sub-family as revealed by the structure of the inhibitor-complexed *Haemophilus influenzae* enzyme. *Journal of Molecular Biology, 303*, 405-421.
97. North, R. A., Kessans, S. A., Atkinson, S. C., Suzuki, H., Watson, A. J., Burgess, B. R., Angley, L. M., Hudson, A. O., Varsani, A., Griffin, M. D., Fairbanks, A. J., & Dobson, R. C. (2013). Cloning, expression, purification, crystallisation and preliminary x-ray diffraction studies of *N*-acetylneuraminate lyase from methicillin-resistant *Staphylococcus aureus*. *Acta Crystallographica Section F, Structural Biology and Crystallisation Communications, 69*, 306-312.
98. Bolt, A., Berry, A., & Nelson, A. (2008). Directed evolution of aldolases for exploitation in synthetic organic chemistry. *Archives of Biochemistry and Biophysics, 474*, 318-330.
99. Iancu, C. V., Zamoon, J., Woo, S. B., Aleshin, A., & Choe, J. Y. (2013). Crystal structure of a glucose/H+ symporter and its mechanism of action. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 17862-17867.
100. Devenish, S. R. A., & Gerrard, J. A. (2009). The quaternary structure of *Escherichia coli N*-acetylneuraminate lyase is essential for functional expression. *Biochemical and Biophysical Research Communications, 388*, 107-111.
101. Kao, C. H., Chen, Y. Y., Wang, L. R., & Lee, Y. C. (2018). Production of *N*-acetyl-D-neuraminic Acid by Recombinant Single Whole Cells Co-expressing *N*-acetyl-D-glucosamine-2-epimerase and *N*-acetyl-D-neuraminic Acid Aldolase. *Molecular Biotechnology, 60*, 427-434.
102. Conejo, M. S., Thompson, S. M., & Miller, B. G. (2010). Evolutionary bases of carbohydrate recognition and substrate discrimination in the ROK protein family. *Journal of Molecular Evolution, 70*, 545-556.
103. Titgemeyer, F., Reizer, J., Reizer, A., & Saier, M. H., Jr. (1994). Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. *Microbiology, 140*, 2349-2354.
104. Martinez, J., Nguyen, L. D., Hinderlich, S., Zimmer, R., Tauberger, E., Reutter, W., Saenger, W., Fan, H., & Moniot, S. (2012). Crystal structures of *N*-acetylmannosamine kinase provide insights into enzyme activity and inhibition. *Journal of Biological Chemistry, 287*, 13656-13665.
105. North, R. A., Seizova, S., Stampfli, A., Kessans, S. A., Suzuki, H., Griffin, M. D., Kvansakul, M., & Dobson, R. C. (2014). Cloning, expression, purification, crystallisation and preliminary x-ray diffraction analysis of *N*-acetylmannosamine kinase from methicillin-resistant *Staphylococcus aureus*. *Acta Crystallographica Section F, Structural Biology and Crystallisation Communications, 70*, 643-649.
106. Caing-Carlsson, R., Goyal, P., Sharma, A., Ghosh, S., Setty, T. G., North, R. A., Friemann, R., & Ramaswamy, S. (2017). Crystal structure of *N*-acetylmannosamine kinase from *Fusobacterium nucleatum*. *Acta Crystallographica F, Structural Biology Communications, 73*, 356-362.
107. Manjunath, L., Guntupalli, S. R., Currie, M. J., North, R. A., Dobson, R. C. J., Nayak, V., & Subramanian, R. (2018). Crystal structures and kinetic analyses of N-acetylmannosamine-6-phosphate 2-epimerases from Fusobacterium nucleatum and Vibrio cholerae. *Acta Crystallographica F, Structural Biology Communications, 74*, 431-440.
108. North, R. A., Kessans, S. A., Griffin, M. D., Watson, A. J., Fairbanks, A. J., & Dobson, R. C. (2014). Cloning, expression, purification, crystallisation and preliminary x-ray diffraction analysis of *N*-acetylmannosamine-6-phosphate 2-epimerase from methicillin-resistant *Staphylococcus aureus*. *Acta Crystallographica F, Structural Biology Communications, 70*, 650-655.
109. Pelissier, M. C., Sebban-Kreuzer, C., Guerlesquin, F., Brannigan, J. A., Bourne, Y., & Vincent, F. (2014). Structural and functional characterisation of the *Clostridium perfringens N*-acetylmannosamine-6-phosphate 2-epimerase essential for the sialic acid salvage pathway. *Journal of Biological Chemistry, 289*, 35215-35224.
110. Allard, S. T., Giraud, M. F., & Naismith, J. H. (2001). Epimerases: structure, function and mechanism. *Cellular and Molecular Life Sciences, 58*, 1650-1665.

111. Samuel, J., & Tanner, M. E. (2002). Mechanistic aspects of enzymatic carbohydrate epimerisation. *Nature Product Reports, 19*, 261-277.
112. Fujiwara, T., Saburi, W., Matsui, H., Mori, H., & Yao, M. (2014). Structural insights into the epimerization of beta-1,4-linked oligosaccharides catalyzed by cellobiose 2-epimerase, the sole enzyme epimerizing non-anomeric hydroxyl groups of unmodified sugars. *Journal of Biological Chemistry, 289*, 3405-3415.
113. Kowatz, T., Morrison, J. P., Tanner, M. E., & Naismith, J. H. (2010). The crystal structure of the Y140F mutant of ADP-L-glycero-D-manno-heptose 6-epimerase bound to ADP-beta-D-mannose suggests a one base mechanism. *Protein Science, 19*, 1337-1343.
114. Park, J. T., & Uehara, T. (2008). How bacteria consume their own exoskeletons (turnover and recycling of cell wall peptidoglycan). *Microbiology and Molecular Biology Reviews, 72*, 211-227, table of contents.
115. Plumbridge, J. (2015). Regulation of the Utilization of Amino Sugars by *Escherichia coli* and *Bacillus subtilis*: Same Genes, Different Control. *Journal of Molecular Microbiology and Biotechnology, 25*, 154-167.
116. Vincent, F., Yates, D., Garman, E., Davies, G. J., & Brannigan, J. A. (2004). The three-dimensional structure of the *N*-acetylglucosamine-6-phosphate deacetylase, NagA, from *Bacillus subtilis*: a member of the urease superfamily. *Journal of Biological Chemistry, 279*, 2809-2816.
117. Ferreira, F. M., Mendoza-Hernandez, G., Castaneda-Bueno, M., Aparicio, R., Fischer, H., Calcagno, M. L., & Oliva, G. (2006). Structural analysis of *N*-acetylglucosamine-6-phosphate deacetylase apoenzyme from *Escherichia coli*. *Journal of Molecular Biology, 359*, 308-321.
118. Ahangar, M. S., Furze, C. M., Guy, C. S., Cooper, C., Maskew, K. S., Graham, B., Cameron, A. D., & Fullam, E. (2018). Structural and functional determination of homologs of the *Mycobacterium tuberculosis N*-acetylglucosamine-6-phosphate deacetylase (NagA). *Journal of Biological Chemistry, 293*, 9770-9783.
119. Vincent, F., Davies, G. J., & Brannigan, J. A. (2005). Structure and kinetics of a monomeric glucosamine 6-phosphate deaminase: missing link of the NagB superfamily? *Journal of Biological Chemistry, 280*, 19649-19655.
120. Broschat, K. O., Gorka, C., Page, J. D., Martin-Berger, C. L., Davies, M. S., Huang Hc, H. C., Gulve, E. A., Salsgiver, W. J., & Kasten, T. P. (2002). Kinetic characterization of human glutamine-fructose-6-phosphate amidotransferase I: potent feedback inhibition by glucosamine 6-phosphate. *Journal of Biological Chemistry, 277*, 14764-14770.
121. Natarajan, K., & Datta, A. (1993). Molecular cloning and analysis of the NAG1 cDNA coding for glucosamine-6-phosphate deaminase from *Candida albicans*. *Journal of Biological Chemistry, 268*, 9206-9214.
122. Calcagno, M., Campos, P. J., Mulliert, G., & Suastegui, J. (1984). Purification, molecular and kinetic properties of glucosamine-6-phosphate isomerase (deaminase) from *Escherichia coli*. *Biochimica et Biophysica Acta, 787*, 165-173.
123. Oliva, G., Fontes, M. R., Garratt, R. C., Altamirano, M. M., Calcagno, M. L., & Horjales, E. (1995). Structure and catalytic mechanism of glucosamine 6-phosphate deaminase from *Escherichia coli* at 2.1 Å resolution. *Structure, 3*, 1323-1332.
124. Liu, C., Li, D., Liang, Y. H., Li, L. F., & Su, X. D. (2008). Ring-opening mechanism revealed by crystal structures of NagB and its ES intermediate complex. *Journal of Molecular Biology, 379*, 73-81.
125. Teplyakov, A., Obmolova, G., Toedt, J., Galperin, M. Y., & Gilliland, G. L. (2005). Crystal Structure of the Bacterial YhcH Protein Indicates a Role in Sialic Acid Catabolism. *Journal of Bacteriology, 187*, 5520-5527.
126. Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R., & Gordon, J. I. (2008). Evolution of mammals and their gut microbes. *Science, 320*, 1647-1651.
127. Jeong, H. G., Oh, M. H., Kim, B. S., Lee, M. Y., Han, H. J., & Choi, S. H. (2009). The capability of catabolic utilisation of *N*-acetylneuraminic acid, a sialic acid, is essential for *Vibrio vulnificus* pathogenesis. *Infection and Immunity, 77*, 3209-3217.
128. Hooper, L. V., Midtvedt, T., & Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annual Review of Nutrition, 22*, 283-307.
129. Brigham, C., Caughlan, R., Gallegos, R., Dallas, M. B., Godoy, V. G., & Malamy, M. H. (2009). Sialic Acid (N-Acetyl Neuraminic Acid) Utilization by *Bacteroides fragilis* Requires a Novel *N*-acetylmannosamine Epimerase. *Journal of Bacteriology, 191*, 3629-3638.
130. Olson, M. E., King, J. M., Yahr, T. L., & Horswill, A. R. (2013). Sialic acid catabolism in *Staphylococcus aureus*. *Journal of Bacteriology, 195*, 1779-1788.
131. Almagro-Moreno, S., & Boyd, E. F. (2010). Bacterial catabolism of nonulosonic (sialic) acid and fitness in the gut. *Gut Microbes, 1*, 45-50.
132. Chubukov, V., Gerosa, L., Kochanowski, K., & Sauer, U. (2014). Coordination of microbial metabolism. *Nature Reviews Microbiology, 12*, 327-340.

133. Deutscher, J. (2008). The mechanisms of carbon catabolite repression in bacteria. *Current Opinion in Microbiology, 11*, 87-93.
134. Gorke, B., & Stulke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Reviews Microbiology, 6*, 613-624.
135. Desvergne, B., Michalik, L., & Wahli, W. (2006). Transcriptional regulation of metabolism. *Physiological Reviews, 86*, 465-514.
136. Plumbridge, J. (2001). DNA binding sites for the Mlc and NagC proteins: Regulation of nagE, encoding the N-acetylglucosamine-specific transporter in *Escherichia coli*. *Nucleic Acids Research, 29*, 506-514.
137. Plumbridge, J. (1995). Co-ordinated regulation of amino sugar biosynthesis and degradation: the NagC repressor acts as both an activator and a repressor for the transcription of the glmUS operon and requires two separated NagC binding sites. *EMBO Journal, 14*, 3958-3965.
138. Alves, R., & Savageau, M. A. (2000). Effect of Overall Feedback Inhibition in Unbranched Biosynthetic Pathways. *Biophysical Journal, 79*, 2290-2304.
139. Hwang, J., Kim, B. S., Jang, S. Y., Lim, J. G., You, D. J., Jung, H. S., Oh, T. K., Lee, J. O., Choi, S. H., & Kim, M. H. (2013). Structural insights into the regulation of sialic acid catabolism by the *Vibrio vulnificus* transcriptional repressor NanR. *Proceedings of the National Academy of Sciences of the United States of America, 110*, E2829-2837.
140. Brennan, R. G., & Matthews, B. W. (1989). The helix-turn-helix DNA-binding motif. *Journal of Biological Chemistry, 264*, 1903-1906.
141. Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., & Pabo, C. O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature, 298*, 447-451.
142. Jain, D. (2015). Allosteric control of transcription in GntR family of transcription regulators: A structural overview. *IUBMB Life, 67*, 556-563.
143. Ptashne, M. (2004). *A genetic switch : phage lambda revisited* (3rd ed.). Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
144. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
145. Johnston, J. W., Zaleski, A., Allen, S., Mootz, J. M., Armbruster, D., Gibson, B. W., Apicella, M. A., & Munson, R. S., Jr. (2007). Regulation of sialic acid transport and catabolism in *Haemophilus influenzae*. *Molecular Microbiology, 66*, 26-39.
146. Gualdi, L., Hayre, J. K., Gerlini, A., Bidossi, A., Colomba, L., Trappetti, C., Pozzi, G., Docquier, J. D., Andrew, P., Ricci, S., & Oggioni, M. R. (2012). Regulation of neuraminidase expression in *Streptococcus pneumoniae*. *BioMed Central Microbiology, 12*, 200.
147. Brennan, R. G. (1993). The winged-helix DNA-binding motif - another helix-turn-helix takeoff. *Cell, 74*, 773-776.
148. Kalivoda, K. A., Steenbergen, S. M., Vimr, E. R., & Plumbridge, J. (2003). Regulation of sialic acid catabolism by the DNA binding protein NanR in *Escherichia coli*. *Journal of Bacteriology, 185*, 4806-4815.
149. Kim, B. S., Hwang, J., Kim, M. H., & Choi, S. H. (2011). Cooperative regulation of the *Vibrio vulnificus* nan gene cluster by NanR protein, cAMP receptor protein, and *N*-acetylmannosamine 6-phosphate. *Journal of Biological Chemistry, 286*, 40889-40899.
150. Beckett, D. (2009). Regulating transcription regulators via allostery and flexibility. *Proceedings of the National Academy of Sciences, 106*, 22035-22036.
151. Fillenberg, S. B., Grau, F. C., Seidel, G., & Muller, Y. A. (2015). Structural insight into operator dre-sites recognition and effector binding in the GntR/HutC transcription regulator NagR. *Nucleic Acids Research, 43*, 1283-1296.
152. Resch, M., Schiltz, E., Titgemeyer, F., & Muller, Y. A. (2010). Insight into the induction mechanism of the GntR/HutC bacterial transcription regulator YvoA. *Nucleic Acids Research, 38*, 2485-2497.

Chapter Two

On the structure and function of the proteins encoded by the *yjhBC* operon – Part One

2.1 Introduction

2.1.1 *yjhBC* operon

The *E. coli* sialoregulon (Figure 2.1) contains three operons that are transcriptionally regulated by the GntR-type NanR from *E. coli*. NanR specifically recognises and binds to tandem repeats of the nucleotide sequence GGTATA within the promoter region of these operons controlling their expression (61; 62).

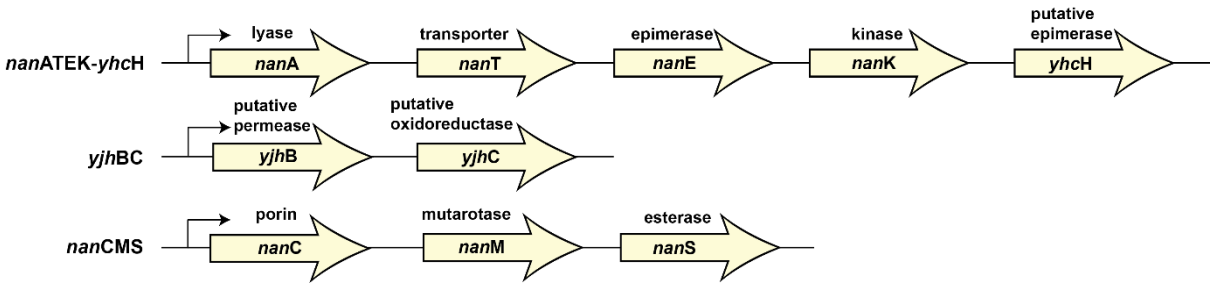


Figure 2.1 | The *Escherichia coli* sialoregulon. In *E. coli*, the transcriptional regulator, NanR negatively controls the expression of the core operon, *nanATEK-yhcH* along with the *nanCMS* operon and uncharacterised *yjhBC* operon.

The *nanATEK-yhcH* operon encodes the core catabolic genes responsible for the import and catabolism of Neu5Ac—the most common form of sialic acid (63; 64). The genes *nanA*, *nanT, nanE* and *nanK* encode a lyase, a sialic acid-specific sugar-proton symporter, an epimerase and a kinase, respectively (64-66). While these enzymes are well characterised in *E. coli*, little is known about the YhcH protein. Although, recent studies within *Haemophilus influenzae*, hypothesise that YhcH functions as an epimerase to process the glycolyl form of sialic acid known as *N*-glycolylneuraminic acid (67); a hydroxylated derivative of Neu5Ac common in most mammals, but not synthesised by humans (68; 69).

The *nanCMS* operon encodes an outer membrane porin (*nanC*), responsible for the selective uptake of sialic acid (70; 71), a sialate mutarotase (*nanM*) that facilitates the conversion of the α-anomer of Neu5Ac to the thermodynamically favourable β-anomer (72), and a 9-O-acetylated Neu5Ac esterase (*nanS*), which permits growth on this sialic acid derivative (73; 74).

Although the biological role of the *E. coli nanATEK-yhcH* and *nanCMS* operons in sialic acid metabolism is reasonably well understood, the biological role of the *yjhBC* operon has yet to be determined.

Bioinformatic analysis places YjhB in the major facilitator superfamily, comparable to the sugar-proton symporter NanT, while YjhC is a member of the Gfo/Idh/MocA family of enzymes. Taken together, the regulation of *yjhBC* by the transcriptional repressor NanR suggests that the biological function of these genes is involved in sialic acid catabolism, where it been hypothesised they function to catabolise less common forms of sialic acid (75). As discussed in Chapter One, Section 1.6, the suite of enzymes that catabolise Neu5Ac are excellent candidates for antimicrobial drug development, targeting pathogenic bacteria who exploit this nutrient source (76-79). Hence, the structural and functional characterisation of these genes will both enhance our understanding of sialic acid catabolism and inform rational drug design to this viable target. YjhC is addressed in this chapter, while YjhB is investigated in Chapter Three.

2.1.2 Gfo/Idh/MocA protein family

Through bioinformatic analysis, YjhC is placed in the glucose-fructose oxidoreductase/inositol dehydrogenase/rhizopine catabolism (Gfo/Idh/MocA) family of enzymes that utilise nicotinamide adenine dinucleotide (NAD) or nicotinamide adenine dinucleotide phosphate (NADP) as a cofactor (Figure 2.2) for an oxidoreductase/ dehydrogenase functionality (80). This cofactor plays a key role during catalysis by either accepting a hydride ion (oxidation reaction) from, or donating a hydride ion (reducing reaction) to the substrate from the C4 position of the nicotinamide ring (80). Typical substrates of the Gfo/Idh/MocA family oxidoreductases are pyranose sugars (81-84). Although, some members show catalytic activity towards other compounds, such as biliverdin; a product of heme catabolism (85) or trans-dihydrodiols; an aromatic hydrocarbon (86). Together, this highlights the diverse substrate specificity within the protein family.

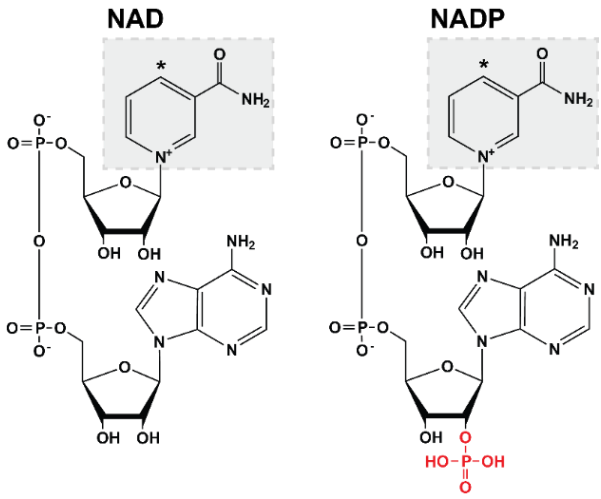


Figure 2.2 | Structures of nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP) cofactors. The additional phosphate group of NADP on the 2' position of the adenine ribose moiety is coloured red. The C4 position of each nicotinamide ring (grey box), which is essential for hydride transfer is highlighted (*).

The typical architecture for this protein family consists of two domains; an N-terminal Rossmann fold (87), which facilitates the binding of NAD/NADP, and a C-terminal α/β-domain believed to share a role in both substrate binding and oligomerisation through a long, predominantly antiparallel β-sheet (Figure 2.3A) (80). The Rossmann fold is one of the most common and widespread structural motifs within the proteome comprising an alternating series of β-strands and α-helices (87; 88). The glycine-rich residues located in a turn between the first β-strand and α-helix can serve as a fingerprint for cofactor specificity, interacting with the pyrophosphate bridge of NAD(P) (Figure 2.3B) While NAD binding enzymes normally possess a GxGxxG consensus sequence after the first β-strand (89), a GxGxxA motif usually correlates with NADP binding by the enzyme (90; 91).

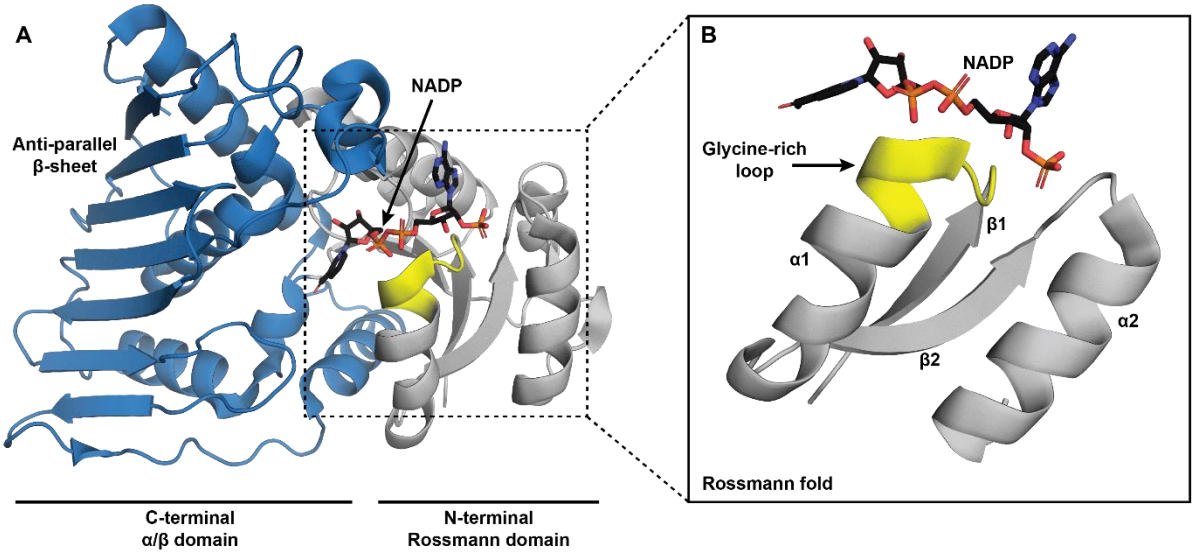


Figure 2.3 | Typical architecture of the Gfo/Idh/MocA protein family and the Rossman fold. A) The N-terminal domain contains the distinctive Rossmann fold which facilitates the binding of the NAD/NADP cofactor (grey). The C-terminal α/β-domain believed to share a role in both substrate binding and oligomerisation (blue). The long, predominantly antiparallel β-sheet is a feature of this protein family. **B)** The β1- α1- β2- α2 segment forms the Rossmann fold. The glycine-rich loop that specifically recognises either NAD or NADP is highlighted in yellow, located between the first β-strand and α-helix (blue) of the Rossmann fold. The GxGxxG motif recognises NAD, while a GxGxxA motif binds NADP. Together, these motifs serve as the fingerprint for cofactor specificity within the Rossmann fold. The monomeric subunit is modelled from the *Caulobacter crescentus* aldose-aldose oxidoreductase (PDB 5A02), a member of the Gfo/Idh/MocA family as a representative structure.

Despite numerous crystal structures of Gfo/Idh/MocA protein family members (defined by the Pfam ID: PF01408) within the PDB, most of these structures are part of structural genomics projects where only the coordinates are reported. Unfortunately, those linked to publications are rarely in complex with their substrate, which therefore restricts the ability to identify key residues that are involved in the

catalytic reaction. However, the consensus motif AGKHVxCE**K**P has been identified as a fingerprint of sugar dehydrogenases (92), where the second lysine residue (in bold) has been suggested to play a fundamental role in the catalytic reaction (82). As an example in the *myo*-inositol dehydrogenase from *Bacillus subtilis*, this lysine residue is involved in a catalytic triad (93), where it assists in positioning the substrate for catalysis and is part of a proton relay system between neighbouring residues. When this lysine residue was mutated to valine, all dehydrogenase activity was abolished, supporting that lysine performs an integral role in the catalytic mechanism (93). Moreover, surrounding residues coordinate the substrate and additionally may play a role in the catalytic mechanism, dependent upon the chemistry (80).

While Gfo/Idh/MocA family members all have similar tertiary structure, their quaternary assembly is seen to vary predominantly between a dimeric and tetrameric assembly (80), although a monomeric structure has been reported (94). Further, the sequence identity between proteins is typically below 20%—a consequence of their diverse functionality. Of the dimeric structures, the most common assembly reported is formed by packing the long, flat-faced β-sheet of the C-terminal domain against the opposing monomer to form a rigid, sandwich-like structure, mediated by face-to-face hydrophobic interactions (Figure 2.4A) (81). Alternatively, a similar interface has been observed between opposing monomers, although another face of the β-sheet is utilised (Figure 2.4B) (83). In the tetrameric structures, two different interface assemblies are also observed. One of these is comparable to the most prevalent dimeric interface, although an additional dimer interacts alongside the other to form an extensive, flat-faced β-sheet in what can be described as a dimer of dimers (Figure 2.4C). The glucose-fructose oxidoreductase from *Zymomonas mobilis* is an example of this tetrameric assembly. Interestingly, this enzyme contains an N-terminal extension, which protrudes into the neighbouring monomer, assisting in the stabilisation of the assembly and coordination of the cofactors adenine moiety—a novel feature of this protein (82). Conversely, the other tetramer forms a weaker, more transient assembly utilising the face of the N-terminal domain as opposed to the long, flat-faced β-sheet within the C-terminal domain (Figure 2.4D) (95).

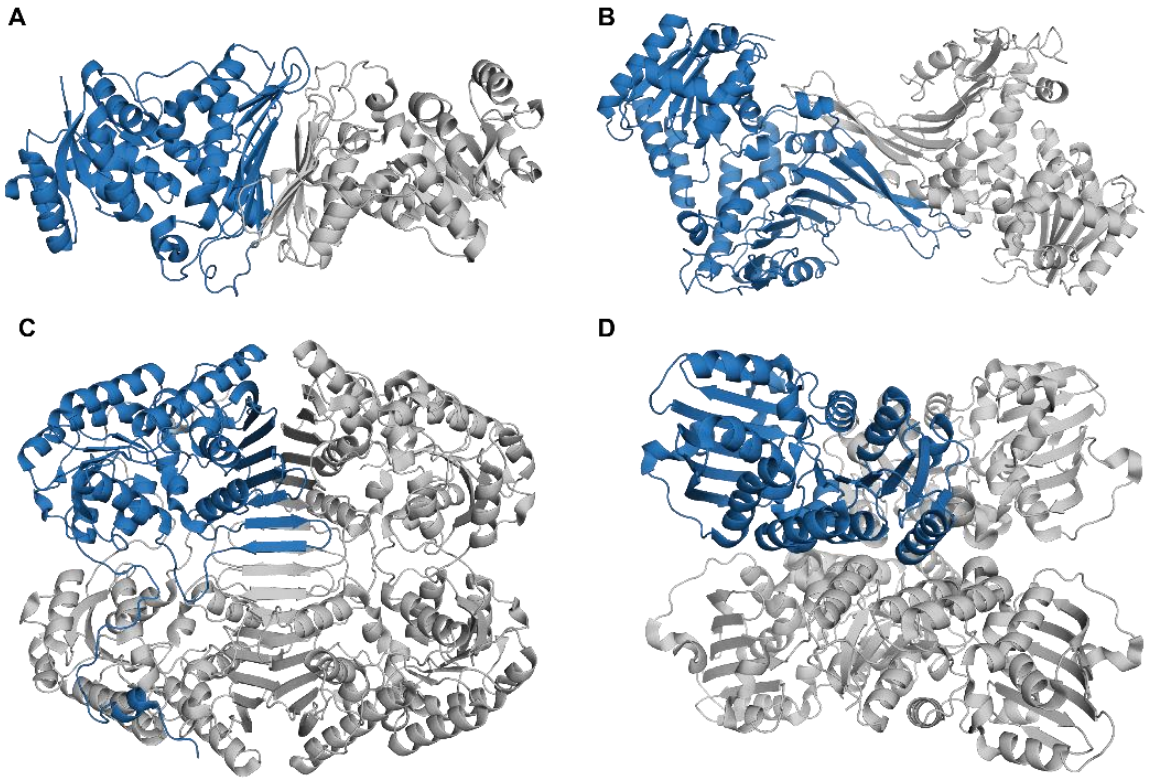


Figure 2.4 | The quaternary structure is observed to vary between Gfo/Idh/MocA family members. A) The most common dimeric assembly is formed by packing the long, flat-faced β-sheet of the C-terminal domain against the opposing monomer to form a rigid, sandwich-like structure, mediated by face-to-face hydrophobic interactions through face (PDB ID 5A02). **B)** This dimeric assembly utilises another face of the β-sheet (PDB ID 1DPG). **C)** Analogous to A, this tetrameric assembly is formed by an additional dimer interacting alongside the other to form an extensive long, flat-faced β-sheet in what can be described as a dimer of dimers. Note the novel N-terminal extension that protrudes to the neighbouring monomer in the glucose-fructose oxidoreductase from *Z. mobilis* (PDB ID 1H6D). **D)** This tetrameric assembly utilises the face of the N-terminal domain as opposed to the long, flat-faced β-sheet within the C-terminal domain (PDB ID 3O9Z).

2.2 Chapter overview

This chapter will investigate the hypothesis that the *yjhBC* operon functions to import and process the less common derivatives of sialic acid by presenting a functional and structure-based approach to address whether YjhC is a *bona fide* oxidoreductase and is truly involved in sialic acid catabolism. Reported as a manuscript, this body of work provides invaluable insight into this uncharacterised enzyme.

2.3 Manuscript

Title:

On the structure and function of *Escherichia coli* YjhC: an oxidoreductase involved in bacterial sialic acid metabolism.

Short title:

On the structure and function of *E. coli* YjhC.

Authors:

Christopher R. Horne [a], Laura Kind [b], James S. Davies [a], Renwick C.J. Dobson [a,c,*]

Author Affiliations:

a) Biomolecular Interaction Centre and School of Biological Sciences, University of Canterbury, PO Box 4800, Christchurch 8041, New Zealand.
b) Department of Biomedicine, University of Bergen, Bergen 5020, Norway.
c) Bio21 Molecular Science and Biotechnology Institute, Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria 3010, Australia.

Acknowledgements:

This work was supported by the following funding to R.C.J.D: 1) the New Zealand Royal Society Marsden Fund (contract UOC1506); 2) a Ministry of Business, Innovation and Employment Smart Ideas grant (contract UOCX1706); and 3) the Biomolecular Interaction Centre (University of Canterbury). This work was also supported by funding from the Canterbury Medical Research Foundation to C.R.H. We acknowledge the Australian Synchrotron, the staff at the MX2 beamline and Dr Nigel Kirby and Dr Tim Ryan for provision of synchrotron facilities along with their assistance in using the beamlines. This research was undertaken in part using the MX2 beamline at the Australian Synchrotron, part of ANSTO, and made use of the ACRF detector. We thank Professor Antony Fairbanks and Dr Stewart Alexander for the chemical synthesis of the sialic acid 1,7 lactone. We thank Dr Rachel North for her insight on bacterial sialic acid metabolism.

Abstract:

Human pathogenic and commensal bacteria evolved the ability to scavenge host-derived sialic acids and subsequently degrade them as a source of nutrition. Expression of the *Escherichia coli yjhBC* operon is controlled by the repressor protein *nanR*, which regulates the core machinery responsible for the import and catabolic processing of sialic acid. The role of the *yjhBC* encoded proteins is not known—here we demonstrate that the enzyme YjhC is an oxidoreductase involved in bacterial sialic acid degradation. First, we demonstrate in vivo using knockout experiments that YjhC is broadly involved in carbohydrate metabolism, including that of *N*-acetyl-D-glucosamine, *N*-acetyl-D-galactosamine and *N*-acetylneuraminic acid. Differential scanning fluorimetry demonstrates that YjhC binds *N*-acetylneuraminic acid and its lactone variant, along with NAD(H), which is consistent with its role as an oxidoreductase. Next, we solved the crystal structure of YjhC in complex with the NAD(H) cofactor to 1.35 Å resolution. The protein fold belongs to the Gfo/Idh/MocA protein family. The dimeric assembly observed in the crystal form is confirmed through solution studies. Ensemble refinement reveals a flexible loop region that may play a key role during catalysis, providing essential contacts to stabilize the substrate—a unique feature to YjhC among closely related structures. Guided by the structure, *in silico* docking experiments support the binding of sialic acid and several common derivatives in the binding pocket with an overall positive charge distribution. Taken together, our results verify the role of YjhC as a *bona fide* oxidoreductase/dehydrogenase and provide the first evidence to support its involvement in sialic acid metabolism.

Keywords:

YjhC, Oxidoreductase, Gfo/Idh/MocA family, NAD, Sialic acid, X-ray Crystallography, Small Angle X-ray Scattering, Analytical Ultracentrifugation

Introduction

Mammalian cell surfaces are decorated with a complex array of glycoconjugates (1; 2). Located at the termini of many cell surface glycoconjugates are sialic acids—a family of nine-carbon amino monosaccharide units (3) that are comprised of more than 50 naturally occurring and structurally distinct variants (4). The most common form of sialic acid is *N*-acetylneuraminic acid, herein abbreviated to Neu5Ac (**Figure 1A**) (3; 4), while other common variants include the lactone (between carbon 1 and 7) and the 9-O-acetylated derivatives of Neu5Ac (**Figure 1B**). Given their terminal location on glycoconjugates, these negatively-charged amino sugars mediate a diverse array of cellular interactions (recognition and adhesion processes), through carbohydrate-binding proteins and their associated receptors (1; 2).

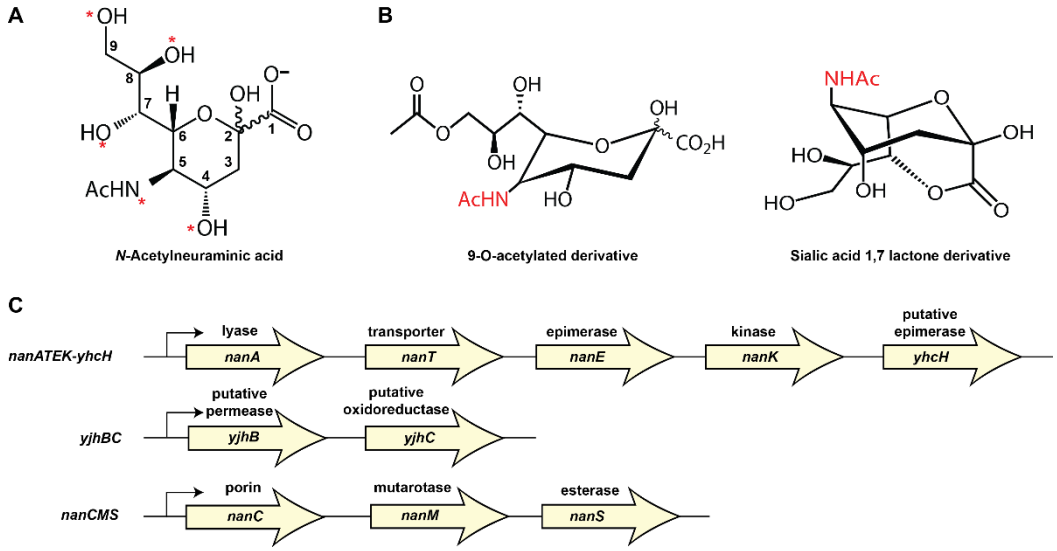


Figure 1 | Structure of common sialic acids and overview of the *Escherichia coli* sialoregulon. A) *N*-Acetylneuraminic acid. The most abundant and best studied of the sialic acids. Carbons within the chemical structure are labelled one through to nine. Substitutions at various hydroxyls (red asterisk) with acetyl moieties or less commonly, phosphate, lactate or succinyl moieties leads to over 50 naturally occurring and structurally distinct derivatives. **B)** The 9-O acetylated derivative and the sialic acid 1,7 lactone derivative are the second and third highest detected sialic acid derivatives in the human gut (5; 61). **C)** Regulation of sialic acid catabolism. In *Escherichia coli*, the transcriptional regulator, NanR negatively controls the expression of the core operon *nanATEK-yhcH*, along with the *nanCMS* and *yjhBC* operons; collectively this is termed the sialoregulon.

Pathogenic and commensal bacteria evolved the ability to scavenge host-derived sialic acids (3), which are a source of carbon, nitrogen and energy (5). In addition, the catabolism of sialic acids yields precursors for peptidoglycan biosynthesis, such as *N*-acetylglucosamine (6; 7). Lastly, some bacteria can decorate their cell surfaces with host-derived sialic acids to camouflage themselves against the host innate immune system (1; 8). Given the accessibility of sialic acids in mucus rich environments, their

utilization offers pathogenic bacteria and commensals, such as *Escherichia coli*, a competitive advantage to colonize and persist within the human host (1; 9; 10).

In *E. coli*, the *nanATEK-yhcH* operon encodes a lyase (*nanA*), a sugar-proton symporter *(nanT*), an epimerase (*nanE*), and a kinase (*nanK*), which together are essential for the utilization of Neu5Ac (5; 6; 8). The expression of these enzymes from the *nanATEK-yhcH* operon is controlled by a transcriptional regulator, NanR, which is in turn regulated by the presence of sialic acids (11). In addition to this core operon, NanR also regulates the *nanCMS* and *yjhBC* operons, collectively termed the sialoregulon (Figure 1C) (11). The *nanCMS* operon encodes an outer membrane porin (*nanC*), which is responsible for the selective import of sialic acid (12; 13), a sialate mutarotase (*nanM*) that catalyses the periplasmic conversion of the α-anomer of Neu5Ac to the thermodynamically favourable β-anomer (14), and a 9-O-acetylated Neu5Ac esterase (*nanS*) that permits growth on this sialic acid derivative (15). Although the biological role of both the *E. coli nanATEK-yhcH* and *nanCMS* operons in sialic acid metabolism is reasonably well understood, the biological role of the *yjhBC* operon has yet to be established—here we focus on the role of *yjhC*.

YjhC is hypothesized to function as an oxidoreductase for less common forms of sialic acid (5). Our bioinformatic analysis places YjhC in the Gfo/Idh/MocA family of oxidoreductase/ dehydrogenase enzymes that use nicotinamide adenine dinucleotide (NAD) or nicotinamide adenine dinucleotide phosphate (NADP) as a cofactor. This analysis was achieved *via* a sequence homology search (see Materials and Methods). Typical substrates of the Gfo/Idh/MocA family oxidoreductases include pyranose sugars (16; 17), although some members have shown catalytic activity towards other substrate types, such as biliverdin; a product of heme catabolism (18) or trans-dihydrodiols; an aromatic hydrocarbon (19).

Here we address two research questions: 1) Is YjhC involved in amino sugar catabolism? and 2) Is YjhC a *bona fide* oxidoreductase/dehydrogenase? To address these questions, we used BiOLOG Phenotype Microarrays (20) to report on the biological role of YjhC, along with thermal shift assays to screen potential substrates for binding. Further, we have solved the crystal structure of *E. coli* YjhC in complex with the cofactor NAD(H) at a resolution of 1.35 Å. Guided by the structure, the substrate-binding pocket was examined, and using *in silico* docking experiments, we support the binding of the acidic sugar, Neu5Ac in its cyclized and open-chain forms, along with the sialic acid 1,7 lactone and the 9-O-acetylated variant.

Materials and Methods

Bioinformatics

To search and compare protein sequences for YjhC, the basic local alignment search tool (*BLAST*) program and *BLASTp* (21) was used. The *DALI* server (22) was used for protein structure comparison. Multiple sequence alignments of YjhC homologues with known atomic structures were generated using *Clustal Omega* (23) and visualized using *ESPript 3.0* (24). Amino acid sequences and atomic structures of homologues were sourced from the UniProt database and Protein Data Bank (PDB), respectively.

Cloning of the YjhC expression construct

The gene encoding YjhC from *E. coli* (UniProt-P39353) was synthesized commercially by GenScript and supplied in a cloning vector, designated pUC57. In order to use a N-terminal His-tag, the *yjhC* gene was digested from pUC57 using the high-fidelity restriction enzymes *Bam*HI and *Hind*III (New England Biolabs), gel purified and ligated with T4 DNA ligase (Takara) to generate pET30ΔSE/*yjhC*. The pET30ΔSE expression vector, conferring kanamycin resistance was kindly supplied by Dr Hironori Suzuki. The recombinant expression vector was transformed into DH5α competent cells, purified using a DNA-Spin™ Plasmid DNA Purification Kit (iNtRon Biotechnology), and verified by DNA sequencing (Macrogen).

Expression and purification of recombinant YjhC

The pET30ΔSE/*yjh*C construct was transformed into *E. coli* BL21(DE3) competent cells (Agilent) and grown in Luria Broth (LB) supplemented with kanamycin (30 mg mL^{-1}) at 37 °C, with shaking at 220 rpm. Expression of YjhC was induced mid-log phase (OD_{600} ≈ 0.6) by the addition of isopropyl β-$_D$-1-thiogalactopyranoside to 1 mM and incubated overnight at 26 °C. Cells were harvested by centrifugation (Thermo Sorvall RC-6-Plus centrifuge) at 8,000 rpm for 10 min and resuspended in buffer A (20 mM Tris−HCl (pH 8.0), 500 mM NaCl, 40 mM imidazole). The cells were lysed by sonication (Hielscher UP200S Ultrasonic Processor) on ice, and cellular debris was pelleted by centrifugation at 16,000 rpm for 30 min.

Purification of YjhC was conducted *via* a two-step procedure: immobilized-metal affinity chromatography (IMAC) and size-exclusion chromatography (SEC). For IMAC the soluble cell lysate was applied onto a His-Trap FF Crude 5 mL column (GE Healthcare) pre-equilibrated with buffer A. The column was washed with buffer A until a steady baseline absorbance was observed. YjhC was eluted using buffer B (20 mM Tris−HCl (pH 8.0), 500 mM NaCl, 400 mM imidazole). Fractions were analyzed by SDS-PAGE and the protein containing fractions were pooled. The sample was spin-concentrated to 500 µL, filtered through a 0.2 µm filter to remove aggregate, and loaded onto a Superdex 200 Increase 10/300 GL SEC column (GE Healthcare) equilibrated with buffer C (20 mM Tris−HCl (pH 8.0), 50 mM

NaCl), for further purification and removal of any aggregates. YjhC was purified to near homogeneity as visualized by SDS-PAGE (Supplementary Figure 1). The absorbance at 280 nm (A_{280}) of the purified protein was measured using a Cary 100-Bio UV-Vis spectrophotometer (Agilent Technologies) and the concentration estimated using a molar extinction coefficient of 42,400 $M^{-1} cm^{-1}$ at 280 nm, as calculated by ProtParam (25). All purification steps were carried out at 4 °C. Purified YjhC was flash-frozen in liquid nitrogen for storage.

BiOLOG Phenotype MicroArray

Commercially available BiOLOG Phenotype MicroArray plates were used to explore the biological role of YjhC. For this experiment, the BL21(DE3) strain was used as the wild-type, while the chromosomal *yjhC* knockout strain, defined as Δ*yjhC*:kan strain was obtained from the *E. coli* Genetic Stock Centre. The 96-well Phenotype MicroArray plates were pre-configured with a variety of compounds and a tetrazolium redox dye, which enabled the phenotypic reaction of growing cells to be monitored in a high-throughput system. Both PM1 and PM2A carbon Phenotype Microarrays were used in this experiment.

For preparation of the inoculum, cells were initially streaked and grown overnight at 37 °C on LB-agar plated media supplemented with 30 mg mL^{-1} kanamycin. After incubation, cells were resuspended in IF-0a inoculating fluid to an OD_{600} of 0.035 using a Cary 100 Bio UV/Vis spectrophotometer (Agilent Technologies). This inoculum was then combined with the tetrazolium redox dye according to the manufacturer's instructions. Carbon sources were rehydrated by adding 100 µL of the inoculum solution to each plate well (position A1 did not have a carbon source and therefore served as a reference/negative control) and incubated aerobically without shaking at 37 °C. Metabolism of the various carbon sources was monitored spectrophotometrically at 570 nm using a SpectraMax M5 plate reader (Molecular Devices) over a 48 h period. The reduction of the tetrazolium dye caused a purple color to develop, which is indicative of active respiration and metabolism of the respective compound. For analysis, the OD_{570} for the reference (position A1) was subtracted from each well and then plotted against each strain to generate a kinetic growth curve. Here, a major phenotypic change between strains was defined as a difference in the rate of respiration by more than 50% after 24 h growth (late log phase), on the same carbon source. The entire experiment was performed in duplicate for both wild-type and Δ*yjhC*:kan bacterial strains. The kinetic growth curves for the PM1 and PM2A plates are presented as supplementary data, along with the contents of each plate (Supplementary Figure 2 and 3). The IF-0a inoculating fluid, tetrazolium dye mix, and carbon utilization plates (PM1 and PM2A) were all purchased from BiOLOG Inc. (Harvard, CA, USA). The raw data is included in Supplementary Figure 4.

Ligand screening using differential scanning fluorimetry

The QuantStudio 3 real-time PCR system (ThermoFisher Scientific) was used to record the thermal stability of the purified protein with and without potential substrates. Reactions consisted of 10 µM protein, 5x SYPRO Orange (prepared as a 50x stock) and the desired concentration of potential substrate (Carbon source = 40 mM, NAD(P) = 5 mM) in a final reaction volume of 25 µL. Samples were heated from 4 °C to a 95 °C at a ramp rate of 0.05 °C, taking fluorescent readings at each time point. Triplicate measurements were performed for each sample. To validate the experiment, one positive control using a well-characterized melting curve (lysozyme at 10 µM) , along with two negative controls (protein only and dye only) were conducted. Data analysis was performed using Protein Thermal Shift software (version 1.3–Applied Biosystems), where an apparent melting point (T_m) of each sample in °C was obtained from the lowest point of the first derivative plot. A summary of all T_m is shown in Supplementary Table 1.

Crystallization, data collection and structure determination of YjhC

Initial crystallization trials for *E. coli* YjhC were performed in-house using the PACT *premier*, Clear Strategy 1, and Clear Strategy 2 commercial screens (Molecular Dimensions). Trials were conducted using the sitting-drop vapor-diffusion method at 20 °C with droplets consisting of 400 nL protein solution and 400 nL reservoir solution. Purified YjhC was concentrated to 12 mg mL^{-1} in buffer C, supplemented with 5 mM of the cofactor NAD, which was identified using differential scanning fluorimetry (Figure 2 and Supplementary Table 1). The crystals were cryo-protected by soaking in 85% reservoir solution and 15% glycerol-ethylene glycol (50:50 mix) prior to being flash-frozen in liquid nitrogen. Condition A1 (1.5 M ammonium sulfate, 0.1 M sodium acetate pH 5.5) from Clear Strategy 2 produced large crystals after four weeks growth. The crystals produced in this condition were suitable for diffraction. Diffraction data were collected (λ = 0.9537 Å) using the MX2 beamline at the Australian Synchrotron with the EIGER X 16 M detector. Data was processed using *XDS* (26) and *AIMLESS* from the CCP4 program suite (27). Details of the data collection statistics are summarized in Table 1.

The structure of YjhC was initially solved using *Auto-Rickshaw* (28) to 1.94 Å resolution, in the spacegroup $P2_12_12_1$. Crystallographic phases were determined by molecular replacement using chain A of PDB entry 5A02; an aldose-aldose oxidoreductase from *Caulobacter crescentus* (17) with a sequence identity of 24%. Diffraction data was collected without the use of cryo-protectant, using litho-loops. The refined output model at 1.94 Å from *Auto-Rickshaw* (28) was used as a search model for molecular replacement to solve the structure of YjhC at 1.35 Å using *PHASER* (29). The model was refined using *REFMAC5* (30) from the CCP4 program suite, and *PHENIX REFINE* (31). To eliminate model bias, simulated annealing using Cartesian restraints was performed. Iterative improvement of the model,

including the addition of ligand and water molecules were performed in *COOT* (32). Water molecules were manually assessed to fit to the observed density and for hydrogen bonding partners. Final refinement was performed using translation-libration-screw (TLS) refinement to produce a final model of *E. coli* YjhC refined to a R_{work} and R_{free} value of 14.98% and 16.87%, respectively. Model validation was performed using *MOLPROBITY* (33). Details of the structural refinement and model validation are summarized in Table 1. To capture possible substrate leads (Neu5Ac and sialic acid 1,7 lactone) in the binding cavity of YjhC, crystals from condition A1 were soaked in cryo-protectant containing 80% reservoir solution, 10% glycerol-ethylene glycol (50:50 mix), and 10% of either Neu5Ac or sialic acid 1,7 lactone, each at a final concentration of 20 mM prior to data collection.

Ensemble refinement

Ensemble refinement was used to evaluate the dynamics and model sources of disorder simultaneously within the final atomic model. Following TLS refinement in *PHENIX REFINE* (31), all missing residues or truncated side chain atoms (due to insufficient electron density) were added using *COOT* (32). Any alternative conformations within the model are automatically removed, while all occupancies are set to one. An ensemble of structures was generated by time-averaged refinement using *PHENIX ENSEMBLE REFINEMENT* (34). The averaging time (τ_x) of structure factors used during the refinement was selected based on the resolution of the diffraction data. Harmonic restraints were applied to all three sulfate ions to prevent stochastic displacement—two were placed at the interface and one closer to the active site. To model the contribution of intramolecular disorder an overall TLS model was applied using one group per chain (34) with a p_{TLS} of 0.8, which defines the number of atoms included in the TLS refinement. A total of 100 ensemble structures were reported where the R_{work} and R_{free} values decreased by 1.9% and 0.9%, respectively. Refinement statistics are presented in Supplementary Table 2.

Analytical ultracentrifugation

Sedimentation velocity experiments were performed with a XL-I analytical ultracentrifuge (Beckman Coulter) using double sector quartz cells and epon center-pieces in an An-50 Ti 8-hole rotor. Data was obtained at 45,000 rpm using 380 µL protein at three different concentrations (7, 14 and 20 µM) in buffer C as the sample, and 400 µL buffer C as the reference. A total of 100 scans were collected at 20 °C using radial absorbance scans at 280 nm and a step size of 0.003 cm. All data were analyzed using *UltraScan 4.0*, release 2578 (35). Sedimentation data were evaluated by the two-dimensional spectrum analysis (2DSA) (36), which affords an unbiased hydrodynamic model, where the sedimentation coefficient and frictional ratio (f/f_0) can be floated independently. Furthermore, 50 Monte-Carlo iterations were performed on the 2DSA data set providing 95% confidence intervals for all measured parameters. All fitting procedures were completed using the *UltraScan* LIMS cluster. The buffer density,

buffer viscosity and an estimate of the partial specific volume of the protein sample based on the amino acid sequence were determined using *UltraScan*. Data analysis are summarized in Table 2.

Small angle X-ray scattering

Small angle X-ray scattering (SAXS) data were collected on the SAXS/WAXS beamline equipped with a Pilatus 1M detector (170 mm × 170 mm, effective pixel size, 172 µm × 172 µm) at the Australian Synchrotron. A sample detector distance of 1600 mm was used, which provided a *q* range of 0.006-0.5 $Å^{-1}$. 50 µL of purified YjhC protein (8 mg mL^{-1}) was injected onto an inline Superdex S200 Increase 5/150 GL SEC column (GE Healthcare), equilibrated with 20 mM Tris–HCl (pH 8.0), 150 mM NaCl, supplemented with 5% (*v*/*v*) glycerol and 0.1% (*w*/*v*) sodium azide, using a flow rate of 0.45 mL min^{-1}. Coflow SAXS was used to minimize sample dilution and maximize signal to noise (37). Scattering data was collected in one second exposures (λ = 1.0332 Å) over a total of 400 frames, using a 1.5 mm glass capillary, at 12 °C. 2D intensity plots were radially averaged, normalized to sample transmission, and background subtracted using the *Scatterbrain* software package (Australian Synchrotron).

Analysis of the scattering data was performed using the *ATSAS* software package (version 2.8.4) (38). *PrimusQT* (39) was used to perform the Guinier analysis and to generate the pairwise distribution function *P*(*r*), calculated using an indirect Fourier transform. Molecular mass was estimated using the *SAXS-MoW2* package (40), while *CRYSOL* (41) was used to evaluate the solution scattering from the known atomic coordinates of YjhC (solved in this study) by fitting to the experimental scattering data. Data collection and analysis are summarized in Table 2.

In silico docking experiments

The molecular interactions that may occur between a small molecule and a protein were predicted using the web service *SwissDock* (42) interfaced with the *CHARMM* package (43).To initiate docking, we defined a search space around the center of mass of our suspected binding cavity as a set of x, y and z coordinates (x center/size = -13.76/17.57; y center/size = 1.11/15.70; z center/size = -19.20/17.85). Secondly, the target protein and ligand were processed to remove the solvent, repair any truncated sidechains using a rotamer library, add hydrogens, assign partial charges and convert all files to a Mol2 file format. Binding modes were then scored using their estimated binding free energy and full fitness, which is defined as the total energy of the system along with a solvation term. All docking results were visualized in *UCSF Chimera* (44) using *ViewDock*. The binding modes are summarized in Supplementary Table 3.

NAD kinetic assay

Kinetic analysis of YjhC was performed using a Cary 100 Bio UV/Vis spectrophotometer (Agilent Technologies), where the reaction was observed by measuring the increase in absorbance at 340 nm due to the consumption of NAD^+. Assays were performed in duplicate, using acrylic cuvettes with a path length of 1 cm. All assay mixtures were equilibrated at 25 °C for 20 min using the heat-controlled Peltier module, prior to measurement. All other stock solutions were kept on ice. The assay mixture consisted of: 20 mM Tris–HCl pH 8.0, 150 mM NaCl, 2 mM NAD^+ (Sigma-Aldrich) and Neu5Ac (Carbosynth), which varied from 5 mM to 60 mM (980 µL total). The assay was initiated following the additional of 20 µL YjhC at a final concentration of 0.05 µM. The assay was repeated using sialic acid 1,7 lactone as the substrate and 0.2 mM NADH (Sigma-Aldrich). In this experiment, the reaction was observed by measuring the decrease in absorption at 340 nm due to the conversion of NADH to NAD^+. Controls without enzyme were performed to measure the background turnover of NAD^+/NADH in solution. All initial rate data were analyzed and fitted with the Michaelis-Menten model using *Prism 8.2.0* (GraphPad Software, La Jolla California USA).

Results and Discussion

YjhC is involved in amino sugar catabolism

A protein's primary sequence is often used to predict its function. However, it is estimated that approximately one-third of all bacterial proteins have such low sequence similarity to do this accurately, meaning many databases annotate protein sequence with an incorrect function (45; 46). This is often the case with homologous proteins that have diverse substrate specificities (47), such as the Gfo/Idh/MocA family. To investigate the function of the oxidoreductase YjhC and its predicted role in sialic acid catabolism, two approaches were taken.

First, we examined the phenotypic consequence of knocking out the *yjhC* gene of *E. coli*, when grown in a range of different carbon sources by using the BiOLOG Phenotype MicroArray system. In this experiment, the overall metabolic activity of the *yjhC* knockout was significantly reduced compared with the wild-type when grown in 15 different carbon sources (Figure 2A). These were L-arabinose, D-arabinose, *N*-acetyl-D-glucosamine, D-trehalose, D-mannose, D-mannitol, D-fructose, α-D-glucose, maltose, α-D-lactose, *N*-acetyl-D-galactosamine, maltotriose, *N*-acetylneuraminic acid, β-D-allose, and D-galactonic acid-γ-lactone. Interestingly, most of these carbon sources are carbohydrates (Figure 2B), where the pyranose sugars displayed the most pronounced catabolic divergence to the wild-type, as identified through the kinetic growth curves (Supplementary Figure 2 and 3). In contrast, an increase in

the metabolic activity of the *yjhC* knockout was observed for four different carbon sources; D-saccharic acid, mucic acid, D- malic acid and L-malic acid, when compared with the wild-type (**Figure 2A**)—the reason for this increase was not clear. Phenotype MicroArrays measure respiration, not growth. As such, where there is reduced metabolism for the *yjh*C knockout compared to the wild-type, the cells may be incapable of converting these sole carbon sources into the downstream metabolites required for cell division, thus leading to potential false negatives (48). Overall these data support the annotation of YjhC within the Gfo/Idh/MocA family, with a pyranose sugar as their likely substrate.

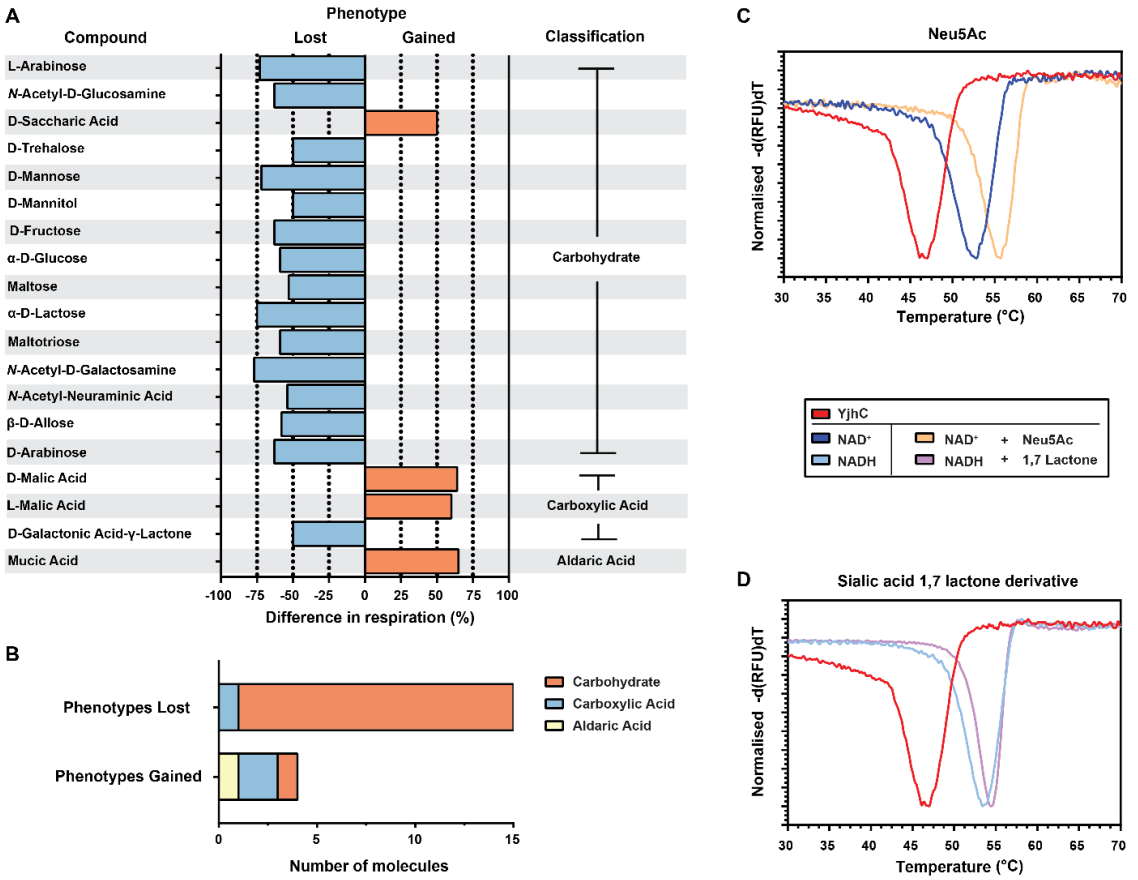


Figure 2 | *In vivo* (BiOLOG Phenotype MicroArray) and *in vitro* (Thermal shift) experiments argue that YjhC is involved in amino sugar metabolism. **A)** The metabolic activity of the *yjhC* gene knockout is significantly reduced (shown in blue) for 15 different carbon sources, when compared with the wild-type. Comparatively, an increase in metabolic activity (shown in orange) of the *yjhC* gene knockout is observed for four different carbon sources. **B)** The carbon sources that significantly affect metabolic activity are presented in categories. Here, carbohydrate-based carbon sources provide the greatest metabolic divergence when the *yjhC* gene is knocked out. **C)** The thermal stability of YjhC in the presence of NAD^+ (52.4 ± 0.2, colored in dark blue) and together with Neu5Ac (55.2 ± 0.1, colored in orange). **D)** The thermal stability of YjhC in the presence of NADH (53.5 ± 0.1, colored in light blue) and together with the sialic acid 1,7 lactone derivative (54.4 ± 0.1, colored in purple). The thermal stability of YjhC without any substrate is 46.5 ± 0.1 (colored in red). The unfolding temperature (T_m) is calculated from the first derivative of the relative fluorescence units (RFU) of the SYPRO Orange dye as a function of temperature (°C). Here, the T_m is denoted as the lowest part of the curve.

Next, we used differential scanning fluorimetry to test whether the compounds implicated in the Phenotype MicroArray experiment bound directly to YjhC. Binding is indicated by a positive or negative thermal shift of the unfolding temperature for the protein (T_m) (49). To verify the role of YjhC as a *bona fide* oxidoreductase, the cofactors NAD^+ and $NADP^+$ were screened. The addition of NAD^+ produced a positive thermal shift of 5.85 °C (Figure 2C), whereas $NADP^+$ had no significant change (0.12 °C). NADH produced a similar shift to the oxidized cofactor of 6.94 °C (Figure 2D). Within the Gfo/Idh/MocA family, members who modify carbohydrates typically catalyze their respective reactions *via* a ping-pong mechanism, where the cofactor binds first, followed by the substrate (16; 17; 50). In the absence of NAD^+/NADH, the carbon compounds tested gave only small thermal shifts, within a range of 0.06-0.37 °C, which we did not consider significant (Supplementary Table 1). However, in the presence of NAD^+/NADH, both Neu5Ac and the sialic acid 1,7 lactone had a significant effect on the thermal stability of YjhC with an additional thermal shift of 2.86 and 0.98 °C, respectively (Figure 2C-D). The remaining carbon compounds did not significantly affect the thermal stability and only produced thermal shifts between 0.01-0.48 °C (Supplementary Table 1).

Taken together, our Phenotype MicroArray and differential scanning fluorimetry experiments suggest that Neu5Ac or analogous sialic acid variants could be the biologically relevant substrate of YjhC. Moreover, we show that YjhC can preferentially bind the cofactor NAD(H) over the cofactor NADP(H), providing strong evidence that YjhC is an oxidoreductase.

YjhC adopts a common oxidoreductase architecture

To add further weight to the hypothesis that YjhC is a *bona fide* oxidoreductase, as predicted from the bioinformatic analysis and thermal shift assays, the structure was solved at a resolution of 1.35 Å (**Table 1**). The structure is in complex with NAD(H) and has two monomers in the asymmetric unit that associate closely to form a dimer (**Figure 3A**), which is discussed later in more detail.

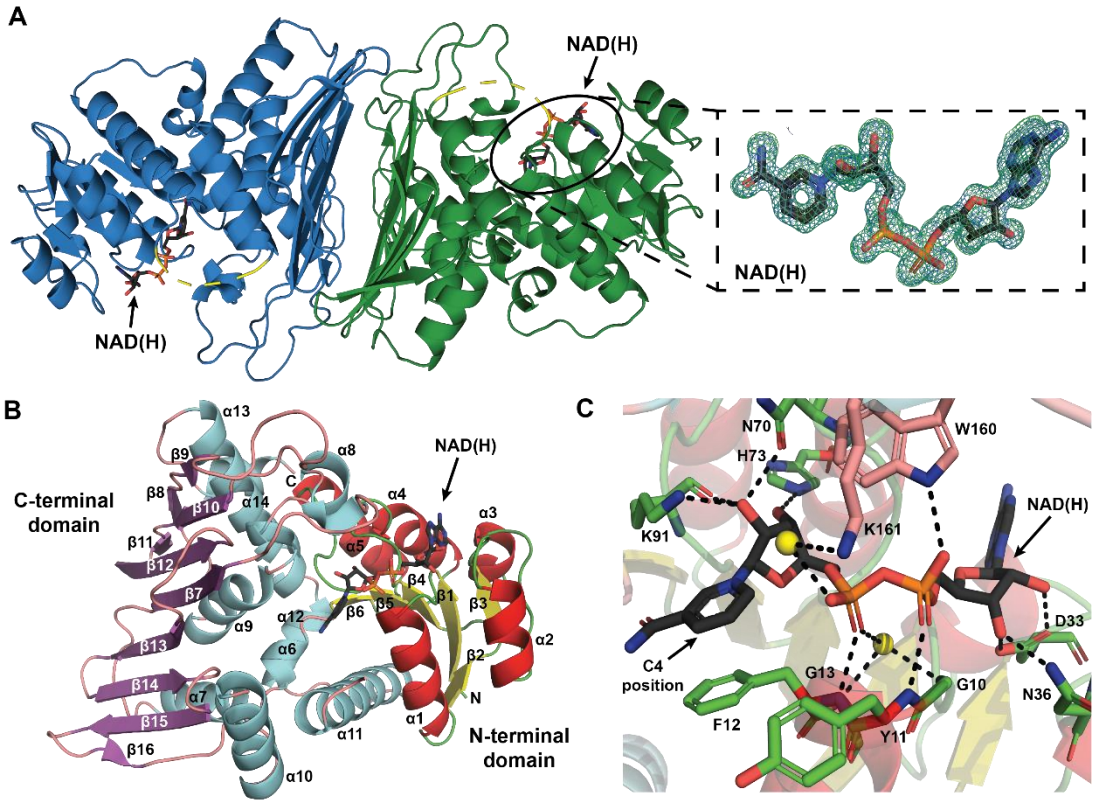


Figure 3 | Structure of *E. coli* YjhC in complex with NAD(H). A) The dimeric assembly present in the asymmetric unit (spacegroup $P\,2_12_12_1$). The disordered region (residues 288-301) is shown as a yellow dotted line. Chains A and B are shown in blue and green, respectively. NAD(H) is shown in a stick representation. The omit map corresponding to NAD(H) in chain A is shown as an insert (black dotted line). Here, the $2F_o$-DF_c electron density map (1.0 σ, blue mesh) and the mF_o-DF_c omit electron density map (3.0 σ, green mesh) is shown. **B)** Monomeric subunit structure of *E. coli* YjhC in complex with NAD(H). Colors are assigned based on secondary structure. The N-terminal domain (residues 1-117) containing the distinctive Rossmann fold is colored yellow, red and green. The C-terminal α/β domain (residues 118-268) is colored magenta, cyan and pink. Alpha (α) helices and beta (β) strands are labelled sequentially. **C)** Close-up view of the active site and NAD(H) binding pocket. Residues within ~3.2 Å of the cofactor are shown (colored as in B) and hydrogen bonding interactions are highlighted by black dashed lines. Ordered water molecules are represented by yellow spheres. The C4 position of the cofactor nicotinamide ring is responsible for hydride transfer to the substrate. All figures were produced using *PyMOL* (62).

Table 1 | Data collection statistics and refinement parameters [a].

Structure	*E. coli* YjhC in complex with NAD(H)
Data collection statistics	
X-ray wavelength (Å)	0.9537
Temperature (K)	100
Detector	EIGER X 16 M (Dectris)
Crystal-to-detector distance (mm)	200
Space group	$P\ 2_1\ 2_1\ 2_1$
a, b, c (Å)	51.51, 84.67, 188.12
α, β, γ (°)	90, 90, 90
Resolution range (Å)	45.18-1.35 (1.398-1.35)
Volume fraction of solvent	50.39
V_m (Å^3/Da)	2.49
$CC_{1/2}$†	0.999 (0.552)
Multiplicity	12.8 (7.8)
Wilson *B*-factor (Å)	18.70
Total No. of reflections	2,315,217 (154,583)
No. of unique reflections	181,008 (17,796)
R_{merge}	0.08 (2.11)
Data completeness (%)	99.0 (99.3)
Average $I/\sigma(I)$	14.2 (1.0)
Refinement statistics	
Resolution (Å)	45.18-1.35 (1.398-1.35)
Reflections used in refinement	180,786 (17,762)
Reflections used for R_{free}	9103 (915)
Total number of residues	710
Total non-H atoms	6629
Protein/ Water	5593/ 933
Ligands	103
Average temperature factors (Å^2)	
Overall	27.48
Protein/ Water	25.76/ 38.49
Ligands	28.18
R_{work} (%)	14.98
R_{free} (%)	16.87
Geometric RMSD	
Bond (Å)	0.008
Angle (°)	1.061
MolProbity statistics ‡	
Rotamer outliers (%)	0.83
Clashscore	2.13
Ramachandran plot	
Favoured/allowed/disallowed regions (%)	97.58/2.42/0.0
PDB Code	6O15

[a] Values in parentheses are for the highest resolution shell

† $CC_{1/2}$ is the correlation between random half-sets of data

‡ *MolProbity* is a structure-validation web service (33)

RMSD = root mean square deviation

The monomer of YjhC shares a typical architecture to other Gfo/Idh/MocA protein family members and consists of two domains (**Figure 3B**). The N-terminal domain (residues 1-117) adopts the distinctive Rossmann fold of six β-strands (51), alongside five neighboring α-helices. The C-terminal α/β domain (residues 118-368) consists of an eight-stranded β-sheet and seven α-helices. This long β-sheet predominately contributes to a flat-faced dimer interface and is a common feature of this protein family (47). The two monomeric chains of YjhC are nearly identical with a root mean square deviation (RMSD) of 0.128 Å (Cα positions).

Within the N-terminal domain, the first residues of a glycine-rich motif in the turn between the first β-strand and α-helix serve as a fingerprint for cofactor specificity, interacting with the pyrophosphate bridge of the cofactor NAD(P). While NADP(H) binding enzymes normally possess a GxGxxA consensus sequence after the first β-strand (52; 53), a GxGxxG motif typically correlates with the binding of NAD(H)—the latter is observed in YjhC between residues G8 and G13. Most residues were well defined in the electron density map, except the sequence $_{288}$SEMDGAIAYGHPGK$_{301}$ within the C-terminal domain, which was disordered in both chains (yellow dotted line in **Figure 3A-B**). Furthermore, three residues at the C-terminus were disordered and thus not modelled in the final refined structure. However, using electrospray ionization mass spectrometry we verified that the full sequence (residues 1-372) was expressed and purified.

Cofactor binding

The crystal structure was solved in complex with the cofactor NAD(H). As shown in **Figure 3A** (insert), electron density for NAD(H) was unambiguously observed for the entire cofactor. NAD(H) was bound within a crevice formed between the loop regions of β-strands, β1 and β4 of the Rossmann fold within the N-terminal domain in which there was a change in the direction of the β-strand order (**Figure 3B**). The binding motif at this region, $_{8}$GxGxxG$_{13}$ confers cofactor specificity for NAD(H) over NADP(H) (54). Specifically, the cofactor was recognized by the side chains of residues F12, D33, N36, N70, H73, K91, W160 and K161, through an interaction that was mediated by an ordered water molecule (**Supplementary Figure 5**). Further, the cofactor was recognized by the main chain atoms of residues G10, Y11 and G13, through an ordered water molecule (**Supplementary Figure 5**). The side chain of residue D33 was observed to form a bifurcated hydrogen bond with the ribose of the adenine moiety. In addition to an increased thermal stability of the protein, which was observed in the presence of NAD^+ over $NADP^+$ (**Supplementary Table 1**), the locality of residue D33 provided further evidence for specificity towards the cofactor NAD(H) as it would sterically hinder the access of the 2'-phosphate group of NADP(H). The oxygen of the primary amide group, located within the nicotinamide moiety, was hydrogen-bonded to the side chain of residues E90, E314, and the main chain carbonyl oxygen of

residue G117, through an ordered water molecule (Supplementary Figure 5). In addition to this extensive hydrogen-bonding network, the nicotinamide ribose moiety was in an *anti*-C2'-endo conformation, which was stabilized through an aromatic edge-to-face interaction, or a C-H-π hydrogen bond with residue F12 $_{(55)}$. The adenine ribose moiety also existed in an *anti*-C2'-endo conformation, and was predominately stabilized through a hydrophobic patch, which was formed by residues Y32, W68, P69, and L72 (Supplementary Figure 5). There was no observed difference between chains A and B with respect to the cofactor binding network. This observation was reinforced in the ensemble structure with residues G10, Y11, F12, G13, D33, N36, N70, H73, K91, and W160, along with NAD(H) itself, displaying little to no movement (Figure 4C). The only observed differences in the ensemble model was residue K161, which displayed increased flexibility relative to neighboring residues, and residue D90, which formed a hydrogen bond with the nitrogen of the amide group of NAD(H). Taken together, the well-ordered cofactor binding network highlights how tightly bound NAD(H) is within the crevice formed at the C-terminal edge of the Rossmann fold. Lastly, this binding mode is analogous to Gfo/Idh/MocA family members who bind NAD(H).

Ensemble refinement demonstrates mobile active site loops

To evaluate the disordered regions of YjhC we combined molecular dynamics simulations with crystallographic data, and generated an ensemble of models by time-averaged refinement, using *PHENIX ENSEMBLE REFINEMENT* (34). This refinement strategy extracts the local molecular motion and samples the global disorder within the protein, through which it can provide more insight into dynamic regions that are functionally important, rather than just an averaged 'snap-shot' of the atomic structure. In this study, the ensemble refinement generated a Boltzmann-weighted population of 100 structures, with a significant reduction of the mF_o-DF_c electron density differences and improved R_{work} and R_{free} factors by 1.9% and 0.9%, respectively (Supplementary Table 2).

Overall, the ensemble model revealed that, while YjhC displayed a well-ordered core, four loop regions (Figure 4A-B) along with the C-terminus displayed varied degrees of flexibility. The four flexible loop regions; L1-L4 are identified through the root mean square fluctuation (RMSF) for the backbone of both chain A and B, which is displayed in Figure 4A. While these loop regions are situated on the surface of the protein, they center around the substrate- and cofactor-binding cavities (Figure 4B). The flexibility in loops L1 and L3 may function to facilitate the diffusion of substrates to the binding cavity, with L3 capping the active site during catalysis—this suggested link between protein flexibility and catalysis has been hypothesized by others (56; 57).

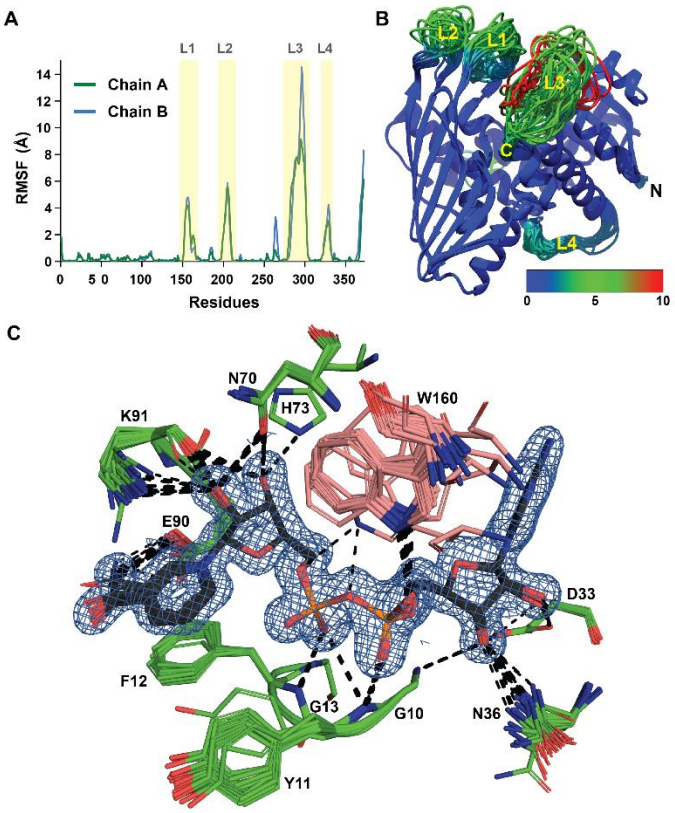


Figure 4 | Ensemble refinement of *E. coli* YjhC. A) Plot displaying the root mean square fluctuation (RMSF) calculated per residue for all backbone atoms in chain A (green) and chain B (blue). **B)** Ensemble structure of the YjhC monomeric subunit colored by RMSF, ranging from 0 to 10 Å (from blue to red). Colored regions, deviating from blue to red highlight regions of high flexibility, which include the N- and C-terminal regions and four flexible loops, spanning residues 150-167 (L1), 198-212 (L2), 288-301 (L3) and 321-335 (L4). **C)** The binding network within the cofactor binding pocket. A representative set of 25 ensemble models are presented, displaying the dynamical fluctuations within the pocket. Key residues (colored as in **Figure 3C**) along with NAD(H) (in black) are shown, highlighting the well-ordered binding network of the cofactor within the pocket. The $2F_o$-DF_c electron density map (blue mesh) is contoured to 1.0 σ. Figures were produced using *PyMOL*(62) and *UCSF Chimera* (44).

E. coli YjhC is a dimer

Within the Gfo/Idh/MocA family the oligomeric assemblies of solved structures are predominately dimeric or tetrameric, although in one case a monomeric structure was reported (58). The structure of *E. coli* YjhC contained two molecules in the asymmetric unit, which form a dimeric assembly (**Figure 3A**), supported by a long, open flat-faced β-sheet in an antiparallel arrangement and *via* a two-fold rotation axis. Using the *PDBePISA* server (59), we determined that approximately 1666 $Å^2$ of accessible surface area was buried against the interface of the opposing monomer, while the total estimated ΔG^{int} value of the dimer interface was approximately -62.3 kcal mol^{-1}. Collectively, this formed a rigid, sandwich-like structure that was stabilized through face-to-face hydrophobic interactions, interdigitation of side chains from the opposing monomers, and a hydrogen-bonding network of nine residues per monomer

(Supplementary Figure 6). In addition, two sulfate ions were modelled at this interface, where were both coordinated to residue R354 of chain A and B. Given their external location and a feature of the crystallization condition, these sulfate ions are unlikely to be of functional relevance.

To assess whether this dimeric assembly in the crystal structure was biologically relevant in solution, we used analytical ultracentrifugation and small angle X-ray scattering. Using analytical ultracentrifugation, three protein concentrations were analyzed (7, 14 and 20 µM) to assess if the dimeric assembly showed any signs of self-association. Following 2DSA-Monte Carlo analysis, all three concentrations were consistent with a single peak and supported a homogenous species in solution (Figure 5A). These data were consistent with the observation in SEC experiments, which showed a single monodisperse species (Supplementary Figure 1). The sedimentation coefficient of 5.60 S (95% confidence interval is 5.49- 5.72) was consistent with a dimeric species, while the frictional ratio (f/f_0) of 1.26 (95% confidence interval is 0.79-1.73), suggested that the molecule was slightly elongated. Further analysis using UltraScan predicted a molecular mass of 88.6 kDa, which was consistent with the calculated dimeric mass of 86.1 kDa (Table 2).

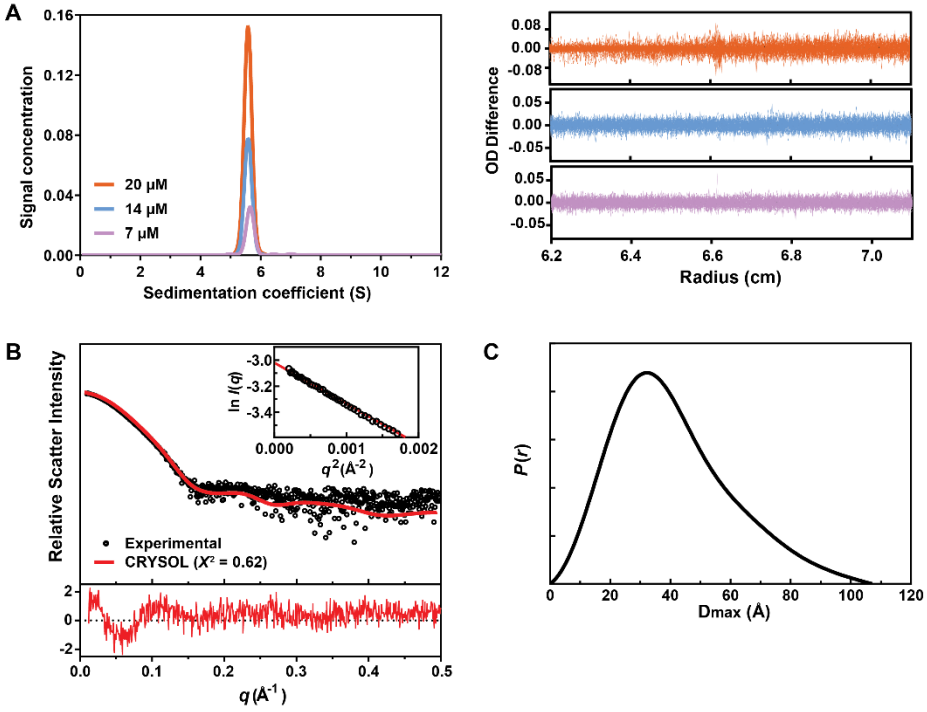


Figure 5 | Solution structure of YjhC. A) 2DSA-Monte Carlo model displaying the sedimentation coefficient (S) distribution of YjhC for three concentrations (7, 14 and 20 µM, colored orange, blue and purple, respectively). The data was best fit to single 5.6 S species, where no sign of self-association is observed. The residuals, shown top right all display randomly distributed noise. **B)** Small angle X-ray scattering data of YjhC is consistent with theoretical scattering of the atomic structure (red line, X^2 value of 0.62). The fit residuals are shown below in red. The Guinier plot shown as an insert, highlights linearity at low q—this provides confidence that the data is free of any significant amount of aggregation or interparticle interference. **C)** Pairwise distribution plot displaying the maximum interparticle dimension (D_{max}) as ~107 Å.

Small angle X-ray scattering data were also consistent with a dimer in solution (**Figure 5B**, SASBDB accession code: SASDFZ3), when analyzed by the online server *SAXS-MoW2* (40). Measured on a relative scale, the molecular weight was 84.5 kDa, which was within 2% of the calculated dimeric mass (86.1 kDa). As shown in **Figure 5B**, the theoretical scattering profile, which was generated from the crystal structure of YjhC, fit the data well to high q, with a reduced X^2 value of 0.62. This supported that the dimeric assembly of YjhC, which was observed in the crystal structure was consistent with our solution structure. Guinier analysis produced a linear plot with a radius of gyration (R_g) of 31.14 Å and a forward scattering value (I_0) of 0.049 (**Table 2**). The linearity of the Guinier plot further supported a single monodisperse species in solution, which was free of any significant amount of aggregation or interparticle interference. An indirect Fourier transform of the scattering data gave a maximum interparticle dimension (D_{max}) of 107.1 Å (**Figure 5C**). This value agreed with the maximum interparticle dimension of ~107 Å, which was determined from the crystal structure. Overall, these data propose *E. coli* YjhC is a tight dimer, where the crystal structure closely models the solution structure.

Table 2 | Summary of data collection and analysis from the solution study of YjhC.

Sedimentation velocity analysis	
Sedimentation coefficient ($\times 10^{-13}$ S)	5.61 (5.51, 5.71)
Frictional ratio (f/f_0)	1.24 (0.97, 1.52)
Molecular weight (kDa)	8.67 (5.96, 11.38)
Partial specific volume of YjhC (mL/g)	0.7329
Buffer density (g/cm^3)	1.0009
Buffer viscosity (cp)	1.0086
SAXS data-collection parameters	
Instrument	Australian Synchrotron SAXS/WAXS beamline
Detector	PILATUS 1M (Dectris)
Wavelength (Å)	1.0332
Maximum flux at sample	8 x 10^{12} photons per second at 12 keV
Camera length (mm)	1600
q range (Å^{-1})	0.006-0.5
Exposure time	Continuous 1 second frame measurements
Sample configuration	SEC-SAXS with co-flow
Sample temperature (°C)	12
SAXS data analysis	
Guinier analysis	
$I(0)$ (cm-1)	0.049 ± 0.00081
R_g (Å)	31.14 ± 0.36
$P(r)$ analysis	
$I(0)$ (cm-1)	0.049 ± 0.00014
R_g (Å)	31.80 ± 0.13
D_{max} (Å)	107.08
Porod volume (Å^{-3})	130000
CRYSOL analysis (X^2 value)	0.62
SAXS-MoW2[a] (Estimated molecular weight (kDa)	84.5

[a] http://saxs.ifsc.usp.br/ (40)

Structural comparison with other Gfo/Idh/MocA family members

A search of the PDB using the *DALI* server (22) was used to gain insight into the overall folding, structural features and possible catalytic functions of YjhC. The numerous search hits were reduced to seven by excluding PDB coordinate entries without a publication available and excluding PDB coordinate entries of identical protein sequence (**Table 3**). A multiple sequence alignment was generated using these seven homologs to investigate consensus regions to and identity characteristic features that may uncover potential catalytic residues (**Supplementary Figure 7**). **Figure 6A** presents the sequence conservation among the search hits mapped onto the monomeric structure of YjhC using *UCSF Chimera*, while **Figure 6B** presents a structural superimposition of the monomeric assembly for each protein. All seven homologs were found to catalyze redox reactions with carbohydrate-based substrates. However, despite sharing a similar overall tertiary fold, the sequence identity between each structure was low compared to YjhC (<25%). This trend was observed across the Gfo/Idh/MocA family, which likely reflects their diverse catalytic function (47).

Table 3 | Result of structural homology search using the *DALI* server.

Protein (*abbreviation*)	Cofactor	Quaternary structure	Z score [a]	r.m.s.d (Å)	% sequence I.D.	PDB ID
UDP-GlucNAcA Dehydrogenase (WlbA)	NAD	Tetramer	37.4	2.3	22	3Q2K [61]
Levoglucosan dehydrogenase (LGDH)	NAD	Tetramer	37.4	2.2	21	6A3I [62]
Scyllo-inositol dehydrogenase (IDH)	NAD	Tetramer	37.1	2.3	19	5YA8 [63]
1,5-anhydro-D-fructose reductase (AFR)	NADP	Dimer	36.4	2.3	23	2GLX [64]
Aldose-aldose oxidoreductase (AAOR)	NADP	Dimer	34.9	2.5	24	5A02 [17]
Glucose-fructose oxidoreductase (GFOR)	NADP	Tetramer	34.4	2.6	24	1H6D [16]
C-3'-ketoreductase (KijD10)	NADP	Tetramer	31.2	2.6	18	3RC2 [65]

[a] Pairwise similarity score measured by the DALI server. Structural homologs are ranked by this score.

The multiple sequence alignment of these structural homologs identified that YjhC contains the sequence $_{83}$NK**KHV**F**CEKP**$_{92}$, which closely resembles the consensus motif AGKHVxCEKP—a fingerprint of sugar dehydrogenases (50). Interestingly, within this motif we identified a conserved *cis*-peptide between the lysine and proline residue for all homologs including YjhC (residues K91 and P92). This *cis* conformation orientates the lysine residue towards the cofactor, where it can form a cation-*pi* interaction with the nicotinamide ring or hydrogen bond to the 2'-hydroxyl group of the nicotinamide ribose. Together, these interactions likely play a crucial role in maintaining the nicotinamide ring in the correct conformation during the catalytic reaction. In addition, this lysine residue is hypothesized to act as a molecular switch, preventing substrate rebinding by moving closer to the nicotinamide ring when the cofactor is reduced (16; 60).

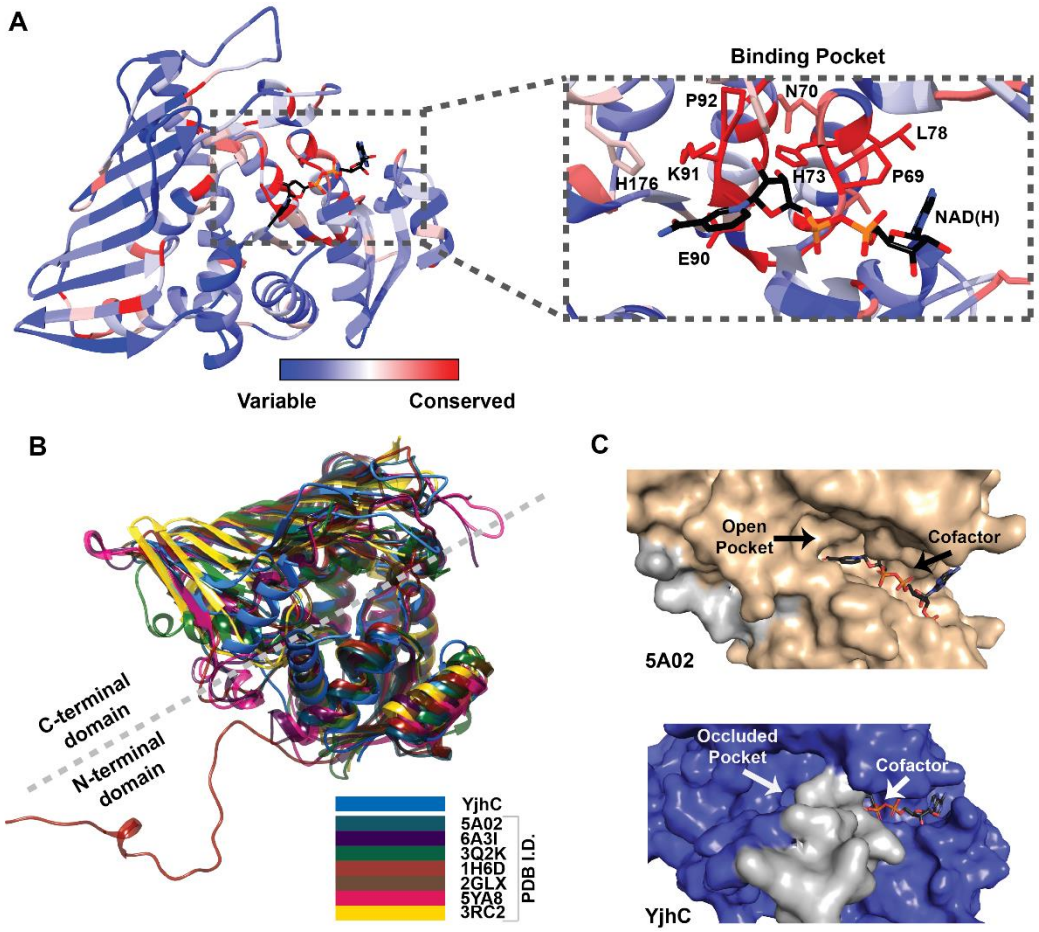


Figure 6 | Structural comparison of Gfo/Idh/MocA family members. A) Mapping of amino acid sequence conservation on YjhC monomeric structure, based on multiple sequence alignment of the seven *DALI*-server search hits in **Table 3**. Depending on the extent of conservation, the residues are colored from red, for the highest conservation, to blue, for variable or non-conserved residues. The insert shows the binding pocket where key residues involved in cofactor binding are well conserved are highlighted. **B)** Structural superimposition of YjhC and all seven *DALI*-server search hits in **Table 3**. Overall the tertiary structure is very similar between enzymes. **C)** Surface representation of the binding pocket of PDB 5A02 (17) (in wheat) and YjhC (in blue). While the pocket is open in PDB 5A02 (similar across all other search hits), the pocket in YjhC is partially occluded by the disordered loop 'L3' (see **Figure 4B**) (in grey). Figures were produced using *UCSF Chimera* (44).

Structural comparison of the binding cavity of these homologous proteins demonstrated that the binding cavity was open (**Figure 6C**), whereas in YjhC the binding cavity was found to be occluded (**Figure 6C**), primarily by disordered loop L3 within the C-terminal domain, which is highly flexible (**Figure 4B**). This loop may have a catalytic role and provide essential contacts to stabilize the substrate—a unique feature of YjhC among closely related structures.

Insight into YjhC substrate specificity at the binding cavity

Electrostatics play a key role in substrate binding, so we mapped the electrostatic potential onto the molecular surface of the substrate binding cavity for *E. coli* YjhC (Supplementary Figure 8A). Situated between the N- and C-terminal domains, this cavity had a positive charge distribution that would support the binding of acidic amino sugars, such as Neu5Ac and its derivatives. This observation was consistent with our data analysis using differential scanning fluorimetry (Figure 2C-D). Typically, within the Gfo/Idh/MocA family the catalyzed reaction precedes *via* a ping-pong mechanism, in which the bound cofactor is required for the substrate to bind. Following binding, the substrate is either oxidized or reduced (47). To initiate the transfer of the hydride ion, the C4 position of the nicotinamide ring must be positioned near the substrate (~3.5 Å), while surrounding residues, depending upon the specific chemistry, can assist in the catalytic reaction.

Interestingly, we observed ambiguous density in the mF_o-DF_c map of the crystal structure that was close to the C4 position of the nicotinamide ring (Supplementary Figure 8B) and close enough to interact with the neighboring K91 and H176 residues. However, no molecule was modelled into this density, as nothing within the crystallization condition appeared to fit the density when refined. Perhaps this could be an artifact from the expression and purification of the protein, with variable occupancy (<1.0).

To help deduce what the unmodelled density was and further identify potential substrates we performed *in silico* docking experiments using *SwissDock* (42), interfaced with the *CHARMM* package (43) on Neu5Ac and several other sialic acid derivatives. These were selected based on their relative distribution in the human gastrointestinal tract, and within the small and large intestine of mice (5). Neu5Ac and its open chain form, along with the 1,7 lactone and 9-O-acetylated sialic acid variant all presented favorable binding modes in which they were: 1) nestled within the binding cavity with favorable electrostatic interactions; 2) positioned close to the C4 position of the nicotinamide ring, poised for hydride transfer; and 3) orientated in a way that a redox reaction would be chemically plausible (Figure 7). These potential ligands were scored based on their estimated ΔG, number of hydrogen bonding partners, and full fitness (Supplementary Table 3), which was defined as the total energy of the system, along with a solvation term (43). As a control, to illustrate the efficiency of the docking strategy, NAD(H) was docked into YjhC with an RMSD difference to the crystal structure of less than 0.5 Å.

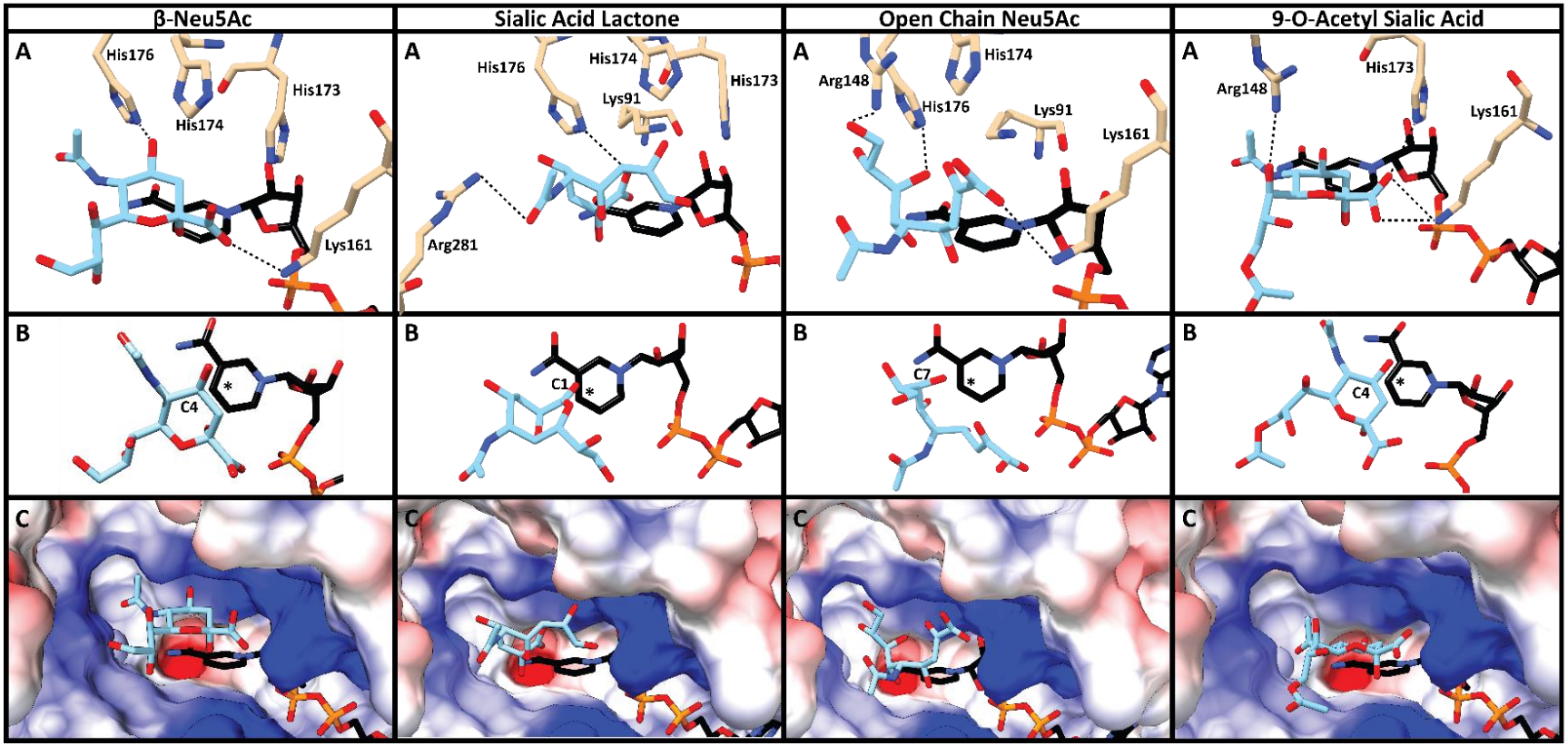


Figure 7 | *In silico* docking experiments of potential substrates. Molecules (light blue sticks) that presented favorable binding modes in proximity to NAD(H) (in black sticks) are presented below. **A)** Key residues within the binding cavity are shown (beige). Hydrogen bonds formed between the substrate and neighboring residues are highlighted (black dash). **B)** Depicts the carbon position of the substrate that is in proximity to the C4 position of the nicotinamide ring. **C)** The substrate is shown in the binding cavity mapped by electrostatic potential using the Adaptive Poisson-Boltzmann Solver (ABPS) plugin. All figures were produced using *UCSF Chimera* (44).

Interestingly, the C4 position of Neu5Ac and the 9-O-acetylated variant were poised for hydride transfer and may suggest the position in which the hydride would be removed (oxidation) or added (reduction) during the catalytic reaction. In contrast, the open chain form of Neu5Ac, and the sialic acid 1,7 lactone variant both presented different carbon positions, relative to the nicotinamide ring. The positioning of the open chain Neu5Ac (C7 position) and the sialic acid 1,7 lactone (C1 position), although not modelled as closely to the C4 position of the nicotinamide could likely be reoriented during conformational changes that inherently may occur during initial binding. In this case, disordered loop L3 (Figure 5B) may be involved in stabilizing the lactone during catalysis. Unfortunately, none of the compounds that were modelled in Supplementary Figure 8, fit the ambiguous density, which supported our hypothesis that this molecule was likely an artifact from the expression and purification of the protein.

Following the identification of several promising leads through docking experiments and differential scanning fluorimetry, attempts were made to co-crystallize the protein with the potential substrates (Neu5Ac and sialic acid 1,7 lactone), or alternatively soak the protein crystals with these molecules in the presence of the cofactor. Unfortunately, following data collection, these was no difference in the electron density within the binding cavity. However, ambiguous density was still observed close to the C4 position of the nicotinamide ring, as illustrated in Supplementary Figure 8B. We believe the presence of this unknown molecule(s) potentially hindered the binding of both Neu5Ac and sialic acid 1,7 lactone.

In addition, we have kinetically tested these promising leads using a spectrophotometric NADH assay, by measuring the change in absorbance at 340 nm upon addition of each molecule. Interestingly, this presented a very small rate for Neu5Ac and gave an apparent K_m of 68.8 mM (Supplementary Figure 9). No observable rate was measured in the absence of YjhC (control), or when using sialic acid 1,7 lactone as the substrate. It is probable that the low observable rate reflects the ambiguous density observed in the binding cavity (Supplementary Figure 8B), and therefore may be inhibiting the reaction in YjhC.

Conclusion

Here, we report the characterization of YjhC from *E. coli* in an effort to elucidate its role in amino sugar metabolism. The results from our *in vivo* (BiOLOG Phenotype MicroArray) and *in vitro* (differential scanning fluorimetry assays) experiments argue that YjhC is involved in sialic acid metabolism. The structure of the protein, the presence of NADH in the active site, and binding studies demonstrate that the enzyme binds NAD(H) as a cofactor, consistent with its role as an oxidoreductase/dehydrogenase from the Gfo/Idh/MocA family. Unfortunately, we were unable to identify the substrate(s) for the enzyme, despite several promising leads, although sialic acid and closely related variants are potential substrates for YjhC.

References

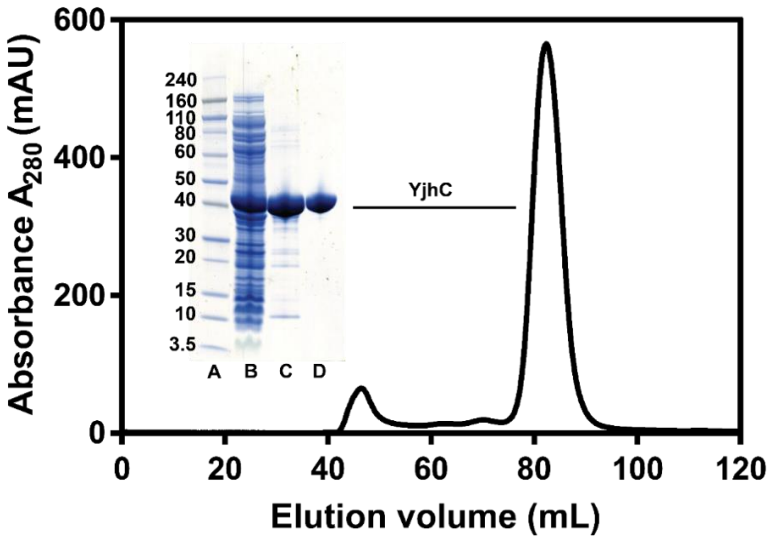
1. Vimr ER, Kalivoda KA, Deszo EL, Steenbergen SM. 2004. Diversity of microbial sialic acid metabolism. Microbiology and Molecular Biology Reviews. 68(1):132-153.
2. Kelm S, Schauer R. 1997. Sialic acids in molecular and cellular interactions. San Diego: Elsevier Academic Press Inc. p. 137-240.
3. Almagro-Moreno S, Boyd EF. 2009. Insights into the evolution of sialic acid catabolism among bacteria. BioMed Central Evolutionary Biology. 9(1):118.
4. Angata T, Varki A. 2002. Chemical diversity in the sialic acids and related alpha-keto acids: An evolutionary perspective. Chemical Reviews. 102(2):439-469.
5. Vimr ER. 2013. Unified theory of bacterial sialometabolism: How and why bacteria metabolise host sialic acids. International Scholarly Research Notices Microbiology. 2013:816713.
6. Severi E, Hood DW, Thomas GH. 2007. Sialic acid utilisation by bacterial pathogens. Microbiology. 153(9):2817-2822.
7. Davies JS, Coombes D, Horne CR, Pearce FG, Friemann R, North RA, Dobson RCJ. 2019. Functional and solution structure studies of amino sugar deacetylase and deaminase enzymes from *Staphylococcus aureus*. FEBS Letters. 593(1):52-66.
8. Vimr ER, Troy FA. 1985. Identification of an inducible catabolic system for sialic acids (*nan*) in *Escherichia coli*. Journal of Bacteriology. 164(2):845-853.
9. Almagro-Moreno S, Boyd EF. 2009. Sialic acid catabolism confers a competitive advantage to pathogenic vibrio cholerae in the mouse intestine. Infection and Immunity. 77(9):3807-3816.
10. Chang DE, Smalley DJ, Tucker DL, Leatham MP, Norris WE, Stevenson SJ, Anderson AB, Grissom JE, Laux DC, Cohen PS et al. 2004. Carbon nutrition of *Escherichia col*i in the mouse intestine. Proceedings of the National Academy of Sciences of the United States of America. 101(19):7427-7432.
11. Kalivoda KA, Steenbergen SM, Vimr ER. 2013. Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. Journal of Bacteriology. 195(20):4689-4701.
12. Condemine G, Berrier C, Plumbridge J, Ghazi A. 2005. Function and expression of an *N*-acetylneuraminic acid-inducible outer membrane channel in *Escherichia coli*. Journal of Bacteriology. 187(6):1959-1965.
13. Wirth C, Condemine G, Boiteux C, Berneche S, Schirmer T, Peneff CM. 2009. NanC crystal structure, a model for outer-membrane channels of the acidic sugar-specific KdgM porin family. Journal of Molecular Biology. 394(4):718-731.
14. Severi E, Müller A, Potts JR, Leech A, Williamson D, Wilson KS, Thomas GH. 2008. Sialic acid mutarotation is catalyzed by the *Escherichia coli* β-propeller protein YjhT. Journal of Biological Chemistry. 283(8):4841-4849.
15. Rangarajan ES, Ruane KM, Proteau A, Schrag JD, Valladares R, Gonzalez CF, Gilbert M, Yakunin AF, Cygler M. 2011. Structural and enzymatic characterization of NanS (YjhS), a 9-O-acetyl N-acetylneuraminic acid esterase from escherichia coli o157:H7. Protein Science. 20(7):1208-1219.
16. Nurizzo D, Halbig D, Sprenger GA, Baker EN. 2001. Crystal structures of the precursor form of glucose-fructose oxidoreductase from *Zymomonas mobilis* and its complexes with bound ligands. Biochemistry. 40(46):13857-13867.
17. Taberman H, Andberg M, Koivula A, Hakulinen N, Penttila M, Rouvinen J, Parkkinen T. 2015. Structure and function of *Caulobacter crescentus* aldose-aldose oxidoreductase. Biochemical Journal. 472:297-307.
18. Kikuchi A, Park SY, Miyatake H, Sun DY, Sato M, Yoshida T, Shiro Y. 2001. Crystal structure of rat biliverdin reductase. Nature Structural Biology. 8(3):221-225.
19. Carbone V, Sumii R, Ishikura S, Asada Y, Hara A, El-Kabbani O. 2008. Structure of monkey dimeric dihydrodiol dehydrogenase in complex with isoascorbic acid. Acta Crystallographica Section D-Biological Crystallography. 64:532-542.
20. Bochner BR, Gadzinski P, Panomitros E. 2001. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. Genome Research. 11(7):1246-1255.
21. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped blast and psi-blast: A new generation of protein database search programs. Nucleic Acids Research. 25(17):3389-3402.
22. Holm L, Rosenstrom P. 2010. DALI server: Conservation mapping in 3d. Nucleic Acids Research. 38(Web Server issue):W545-549.
23. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soeding J et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Molecular Systems Biology. 7.

24. Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new endscript server. Nucleic Acids Research. 42(W1):W320-W324.
25. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. Expasy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research. 31(13):3784-3788.
26. Kabsch W. 2010. XDS. Acta Crystallographica Section D: Biological Crystallography. 66(2):125-132.
27. Evans PR, Murshudov GN. 2013. How good are my data and what is the resolution? Acta Crystallographica Section D-Biological Crystallography. 69:1204-1214.
28. Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS, Tucker PA. 2005. Auto-Rickshaw: An automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. Acta Crystallographica Section D. 61(4):449-457.
29. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. 2007. Phaser crystallographic software. Journal of Applied Crystallography. 40(Pt 4):658-674.
30. Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA. 2011. Refmac5 for the refinement of macromolecular crystal structures. Acta Crystallographica Section D, Biological Crystallography. 67(Pt 4):355-367.
31. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD. 2012. Towards automated crystallographic structure refinement with Phenix.Refine. Acta Crystallographica Section D-Structural Biology. 68:352-367.
32. Emsley P, Cowtan K. 2004. Coot: Model building tools for molecular graphics. Acta Crystallographica Section D, Biological Crystallography. 60(Pt 12 Pt 1):2126-2132.
33. Chen VB, Arendall WB, III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. 2010. Molprobity: All-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D. 66:12-21.
34. Burnley BT, Afonine PV, Adams PD, Gros P. 2012. Modelling dynamics in protein crystal structures by ensemble refinement. eLife. 1:29.
35. Demeler B. 2005. Ultrascan - a comprehensive data analysis software package for analytical ultracentrifugation experiments. In: Scott DJ, Harding SE, Rowe AJ, editors. Analytical ultracentrifugation: Techniques and methods. The Royal Society of Chemistry. p. 210-230.
36. Brookes E, Cao WM, Demeler B. 2010. A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. European Biophysical Journal Biophysical Letters. 39(3):405-414.
37. Kirby N, Cowieson N, Hawley AM, Mudie ST, McGillivray DJ, Kusel M, Samardzic-Boban V, Ryan TM. 2016. Improved radiation dose efficiency in solution SAXS using a sheath flow sample environment. Acta Crystallographica Section D-Structural Biology. 72:1254-1266.
38. Franke D, Petoukhov MV, Konarev PV, Panjkovich A, Tuukkanen A, Mertens HDT, Kikhney AG, Hajizadeh NR, Franklin JM, Jeffries CM et al. 2017. Atsas 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. Journal of Applied Crystallography. 50:1212-1225.
39. Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, Svergun DI. 2003. Primus: A windows pc-based system for small-angle scattering data analysis. Journal of Applied Crystallography. 36:1277-1282.
40. Fischer H, Neto MD, Napolitano HB, Polikarpov I, Craievich AF. 2010. Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. Journal of Applied Crystallography. 43:101-109.
41. Svergun D, Barberato C, Koch MH. 1995. Crysol - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. Journal of Applied Crystallography. 28(6):768-773.
42. Grosdidier A, Zoete V, Michielin O. 2011. Swissdock, a protein-small molecule docking web service based on eadock dss. Nucleic Acids Research. 39:W270-W277.
43. Grosdidier A, Zoete V, Michielin O. 2011. Fast docking using the CHARMM force field with EADock dss. Journal of Computational Chemistry. 32(10):2149-2159.
44. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera - a visualization system for exploratory research and analysis. Journal of Computational Chemistry. 25(13):1605-1612.
45. Chang YC, Hu ZJ, Rachlin J, Anton BP, Kasif S, Roberts RJ, Steffen M. 2016. Combrex-db: An experiment centered database of protein function: Knowledge, predictions and knowledge gaps. Nucleic Acids Research. 44(D1):D330-D335.
46. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. PLoS Computational Biology. 5(12):13.

47. Taberman H, Parkkinen T, Rouvinen J. 2016. Structural and functional features of the NAD(P) dependent gfo/idh/moca protein family oxidoreductases. Protein Science. 25(4):778-786.
48. Bochner BR. 2009. Global phenotypic characterization of bacteria. FEMS microbiology reviews. 33(1):191-205.
49. Niesen FH, Berglund H, Vedadi M. 2007. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nature Protocols. 2(9):2212-2221.
50. Wiegert T, Sahm H, Sprenger GA. 1997. Expression of the *Zymomonas mobilis gfo* gene for NADP-containing glucose:Fructose oxidoreductase (GFOR) in *Escherichia coli* - formation of enzymatically active pregfor but lack of processing into a stable periplasmic protein. European Journal of Biochemistry. 244(1):107-112.
51. Rao ST, Rossmann MG. 1973. Comparison of super-secondary structures in proteins. Journal of Molecular Biology. 76(2):241.
52. Hanukoglu I, Gutfinger T. 1989. Cdna sequence of adrenodoxin reductase - identification of NADP-binding sites in oxidoreductases. European Journal of Biochemistry. 180(2):479-484.
53. Hua YH, Wu CY, Sargsyan K, Lim C. 2014. Sequence-motif detection of NAD(P)-binding proteins: Discovery of a unique antibacterial drug target. Scientific Reports. 4:7.
54. Bellamacina CR. 1996. The nicotinamide dinucleotide binding motif: A comparison of nucleotide binding proteins. FASEB Journal. 10(11):1257-1269.
55. Nishio M, Umezawa Y, Fantini J, Weiss MS, Chakrabarti P. 2014. Ch–π hydrogen bonds in biological macromolecules. Physical Chemistry Chemical Physics. 16(25):12648-12683.
56. Garcia-Viloca M, Gao J, Karplus M, Truhlar DG. 2004. How enzymes work: Analysis by modern rate theory and computer simulations. Science. 303(5655):186-195.
57. Ramanathan A, Agarwal PK. 2011. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biology. 9(11):e1001193.
58. Whitby FG, Phillips JD, Hill CP, McCoubrey W, Maines MD. 2002. Crystal structure of a biliverdin ix alpha reductase enzyme-cofactor complex. Journal of Molecular Biology. 319(5):1199-1210.
59. Krissinel E, Henrick K. 2005. Detection of protein assemblies in crystals. In: Berthold MR, Glen R, Diederichs K, Kohlbacher O, Fischer I, editors. Computational life sciences, proceedings. p. 163-174.
60. Schu M, Faust A, Stosik B, Kohring GW, Giffhorn F, Scheidig AJ. 2013. The structure of substrate-free 1,5-anhydro-d-fructose reductase from *Sinorhizobium meliloti* 1021 reveals an open enzyme conformation. Acta Crystallographica Section F Structural Biology. 69(Pt 8):844-849.
61. Thoden JB, Holden HM. 2011. Biochemical and structural characterization of WlbA from *Bordetella pertussis* and *Chromobacterium violaceum*: Enzymes required for the biosynthesis of 2,3-diacetamido-2,3-dideoxy-d-mannuronic acid. Biochemistry. 50(9):1483-1491.
62. Sugiura M, Nakahara M, Yamada C, Arakawa T, Kitaoka M, Fushinobu S. 2018. Identification, functional characterization, and crystal structure determination of bacterial levoglucosan dehydrogenase. Journal of Biological Chemistry. 293(45):17375-17386.
63. Fukano K, Ozawa K, Kokubu M, Shimizu T, Ito S, Sasaki Y, Nakamura A, Yajima S. 2018. Structural basis of l-glucose oxidation by scyllo-inositol dehydrogenase: Implications for a novel enzyme subfamily classification. PloS One. 13(5).
64. Dambe TR, Kuhn AM, Brossette T, Giffhorn F, Scheidig AJ. 2006. Crystal structure of NADP(H)-dependent 1,5-anhydro-D-fructose reductase from sinorhizobium morelense at 2.2 Å resolution: Construction of a nadh-accepting mutant and its application in rare sugar synthesis. Biochemistry. 45(33):10030-10042.
65. Kubiak RL, Holden HM. 2011. Combined structural and functional investigation of a C-3''-ketoreductase involved in the biosynthesis of dTDP-L-Digitoxose. Biochemistry. 50(26):5905-5917.
66. North RA, Horne CR, Davies JS, Remus DM, Muscroft-Taylor AC, Goyal P, Wahlgren WY, Ramaswamy S, Friemann R, Dobson RCJ. 2018. "Just a spoonful of sugar...": Import of sialic acid across bacterial cell membranes. Biophysical Reviews. 10(2):219-227.
67. DeLano WL. 2002. The PyMOL molecular graphics program. San Carlos: Delano Scientific.

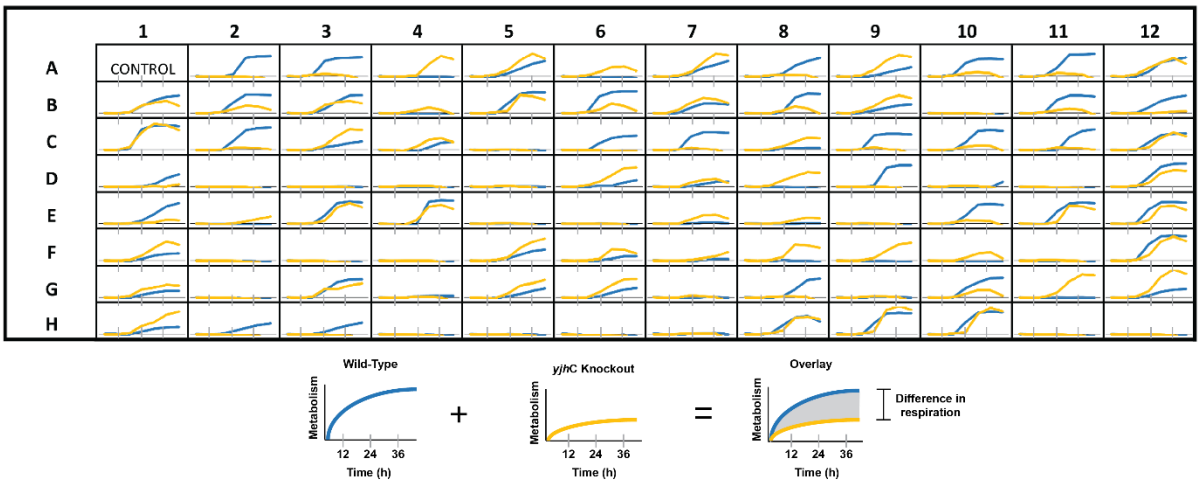
Supplementary Information

Supplementary Figure 1 | YjhC purification. Size exclusion trace of YjhC highlights a homogeneous, monodisperse species. Shown as an insert - SDS-PAGE gel of YjhC purification. **A)** Lane 1: Novex Sharp Pre-Stained Protein Standard ladder (molecular weights are shown in kDa). **B)** Lane 2: cell free supernatant. **C)** Lane 3: pooled fractions following elution in immobilised nickel ion His-trap chromatography. **D)** Lane 4: pooled fraction following size exclusion chromatography. YjhC is present at approximately 40 kDa.



Supplementary Figure 2 | PM1 Phenotype Microarray. Kinetic growth curves for the wild-type (blue) and *yjhC* gene knockout (orange) strains. Each curve contains plots of metabolism (due to dye reduction) against time (0-48 h) for each well and strain. The well contents are presented below.

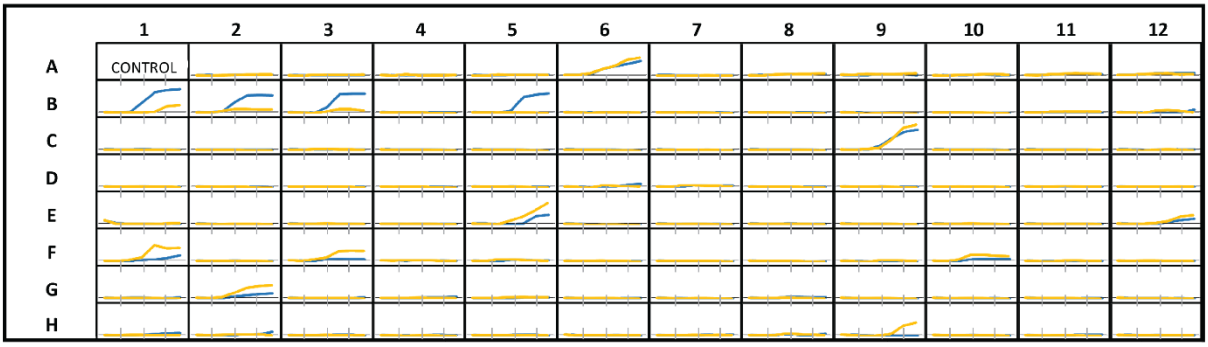
PM1 BIOLOG Microplate

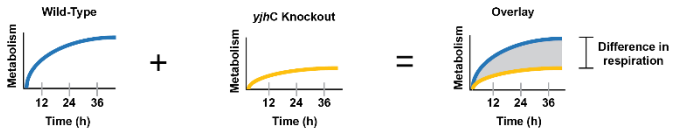


A01 Negative Control	**A02** L-Arabinose	**A02** *N*-Acetyl-D-Glucosamione	**A04** D-Saccharic Acid	**A05** Succinic Acid	**A06** D-Galactose	**A07** L-Aspartic Acid	**A08** L-Proline	**A09** D-Alanine	**A10** D-Trehalose	**A11** D-Mannose	**A12** Dulcitol
B01 D-Serine	**B02** D-Sorbitol	**B03** Glycerol	**B04** L-Frucose	**B05** D-Glucuronic Acid	**B06** D-Gluconic Acid	**B07** D,L-α-Glycerol Phosphate	**B08** D-Xylose	**B09** L-Lactic Acid	**B10** Formic Acid	**B11** D-Mannitol	**B12** L-Glutamic Acid
C01 D-Glucose-6-Phosphate	**C02** D-Galactonic Acid-γ-Lactone	**C03** D,L-Malic Acid	**C04** D-Ribose	**C05** Tween 20	**C06** L-Rhamnose	**C07** D-Fructose	**C08** Acetic Acid	**C09** α-D-Glucose	**C10** Maltose	**C11** D-Melibiose	**C12** Thymidine
D01 L-Asparagine	**D02** D-Aspartic Acid	**D03** D-Glucosaminic Acid	**D04** 1,2-Propanediol	**D05** Tween 40	**D06** α-Keto-Glutaric Acid	**D07** α-Keto-Butyric Acid	**D08** α-Methyl-D-Galactoside	**D09** α-D-Lactose	**D10** Lactulose	**D11** Sucrose	**D12** Uridine
E01 L-Glutamine	**E02** m-Tartaric Acid	**E03** D-Glucose-1-Phosphate	**E04** D-Fructose-6-Phosphate	**E05** Tween 80	**E06** α-Hydroxy-Glutaric Acid -γ-Lactone	**E07** α-Hydroxy Butyric Acid	**E08** β-Methyl-D-Glucoside	**E09** Adonitol	**E10** Maltotriose	**E11** 2-Deoxy Adenosine	**E12** Adenosine
F01 Glycyl-L-Aspartic Acid	**F02** Citric Acid	**F03** m-Inositol	**F04** D-Threonine	**F05** Fumaric Acid	**F06** Bromo Succinic Acid	**F07** Propionic Acid	**F08** Mucic Acid	**F09** Glycolic Acid	**F10** Glyoxylic Acid	**F11** D-Cellobiose	**F12** Inosine
G01 Glycyl-L-Glutamic Acid	**G02** Tricarballylic Acid	**G03** L-Serine	**G04** L-Threonine	**G05** L-Alanine	**G06** L-Alanyl-Glycine	**G07** Acetoacetic Acid	**G08** *N*-Acetyl-β-D-Mannosamine	**G09** Mono Methyl Succinate	**G10** Methyl Pyruvate	**G11** D-Malic Acid	**G12** L-Malic Acid
H01 Glycyl-L-Proline	**H02** p-Hydroxy Phenyl Acetic Acid	**H03** m-Hydroxy Phenyl Acetic Acid	**H04** Tyramine	**H05** D-Psicose	**H06** L-Lyxose	**H07** Glucuronamide	**H08** Pyruvic Acid	**H09** L-Galactonic Acid-γ-Lactone	**H10** D-Galacturonic Acid	**H11** Phenyleth-yl-amine	**H12** 2-Aminoethanol

Supplementary Figure 3 | PM2A Phenotype Microarray. Kinetic growth curves for the wild-type (blue) and *yjhC* gene knockout (orange) strains. Each curve contains plots of metabolism (due to dye reduction) against time (0-48 h) for each well and strain. The well contents are presented below.

PM2A BIOLOG Microplate

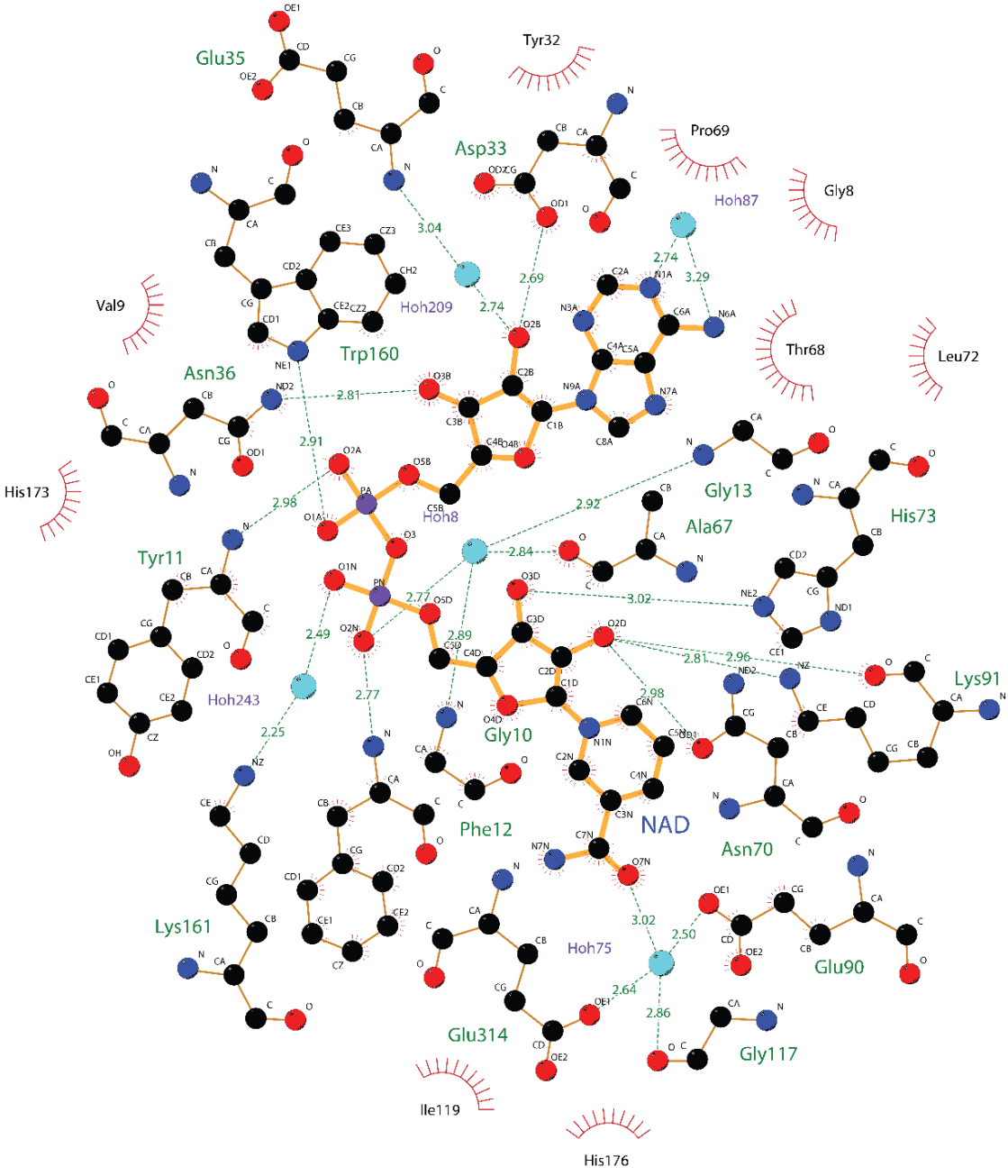




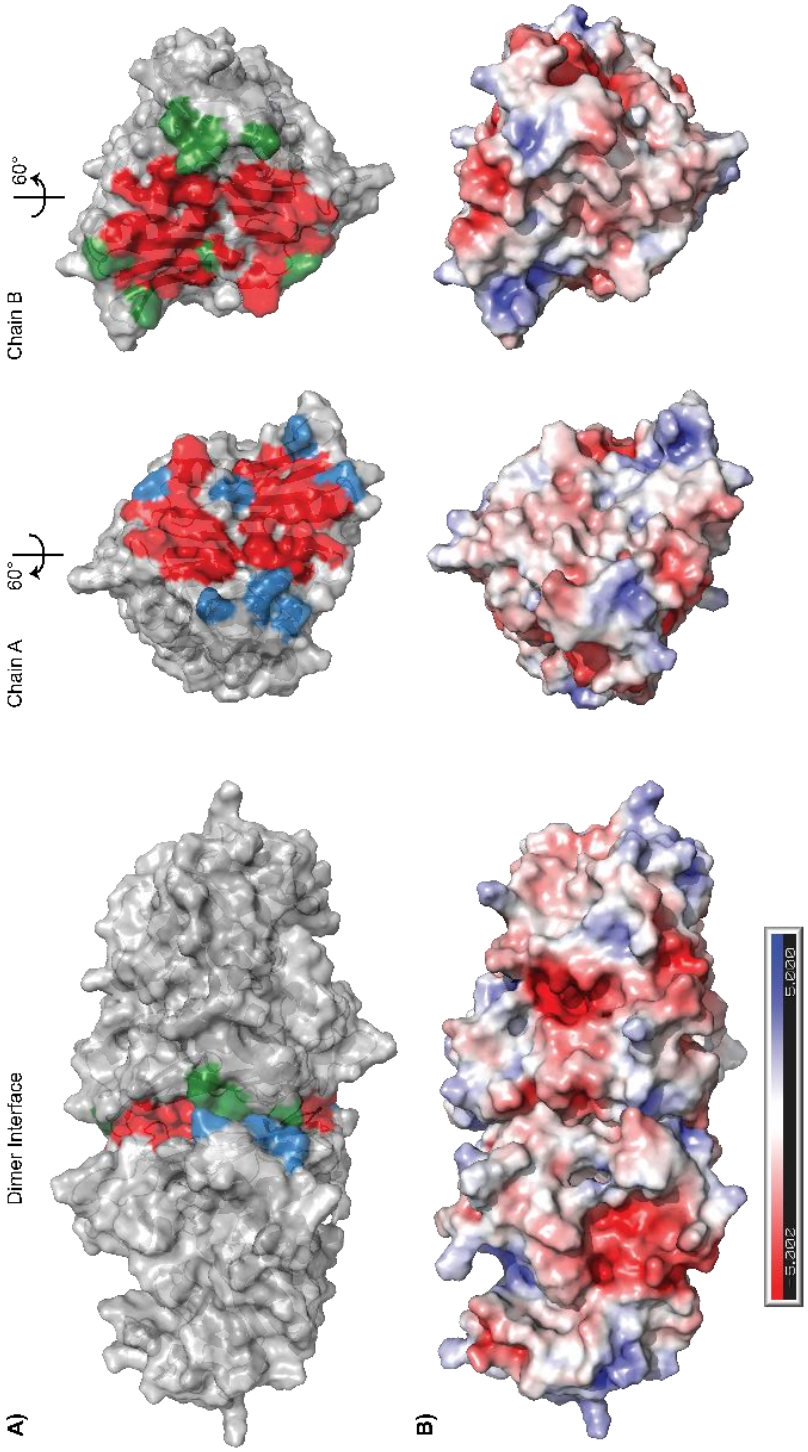
A01 Negative Control	**A02** Chondroitin Sulfate	**A02** α-Cyclodextrin	**A04** β-Cyclodextrin	**A05** γ-Cyclodextrin	**A06** Dextrin	**A07** Gelatin	**A08** Glycogen	**A09** Inulin	**A10** Laminarin	**A11** Mannan	**A12** Pectin
B01 *N*-Acetyl-D-galactosamine	**B02** *N*-Acetyl-neuraminic acid	**B03** β-D-Allose	**B04** Amygdalin	**B05** D-Arabinose	**B06** D-Arabitol	**B07** L-Arabitol	**B08** Arbutin	**B09** 2-Deoxy-D-Ribose	**B10** i-Erythritiol	**B11** D-Fructose	**B12** 3-0-β-D-Galacto-pyranosyl -D-Arabinose
C01 Gentiobiose	**C02** L-Glucose	**C03** Lactitol	**C04** D-Melezitose	**C05** Maltitol	**C06** a-Methyl-D-Glucoside	**C07** a-Methyl-D-Glucoside	**C08** 3-Methyl Glucose	**C09** β-Methyl-D-Glucuronic Acid	**C10** α-Methyl-D-Mannoside	**C11** β-Methyl-D-Xyloside	**C12** Palatinose
D01 D-Raffinose	**D02** Salicin	**D03** Sedoheptulosan	**D04** L-Sorbose	**D05** Stachyose	**D06** D-Tagtose	**D07** Turanose	**D08** Xylitol	**D09** *N*-Acetyl-D-Glucosaminitol	**D10** γ-Amino Butyric Acid	**D11** δ-Amino Valeric Acid	**D12** Butyric Acid
E01 Capric Acid	**E02** Caproic Acid	**E03** Citraconic Acid	**E04** Citramalic Acid	**E05** D-Glucosamine	**E06** 2-Hydroxy Benzoic Acid	**E07** 4-Hydroxy Benzoic Acid	**E08** β-Hydroxy Butyric Acid	**E09** γ-Hydroxy Butyric Acid	**E10** a-Keto-Valeric Acid	**E11** Itaconic Acid	**E12** 5-Keto-D-Glu-conic Acid
F01 D-Lactic Acid Methyl Ester	**F02** Malonic Acid	**F03** Melibionic Acid	**F04** Oxalic Acid	**F05** Oxalomalic Acid	**F06** Quinic Acid	**F07** D-Ribose-1,4-Lactone	**F08** Sebacic Acid	**F09** Sorbic Acid	**F10** Succinamic Acid	**F11** D-Tartaric Acid	**F12** L-Tartaric Acid
G01 Acetamide	**G02** L-Alaninamide	**G03** *N*-Acetyl-L-Glutamic Acid	**G04** L-Arginine	**G05** Glycine	**G06** L-Histidine	**G07** L-Homoserine	**G08** Hydroxyl-L-Proline	**G09** L-Isoleucine	**G10** L-Leicine	**G11** L-Lysine	**G12** L-Methionine
H01 L-Ornithine	**H02** L-Phenylalanine	**H03** L-Pyroglutamic Acid	**H04** L-Valine	**H05** D,L-Carnitine	**H06** Sec-Butyl-amine	**H07** D,L-Octopamine	**H08** Putrescine	**H09** Dihdroxy Acetone	**H10** 2,3-Butanediol	**H11** 2,3-Butanedi-one	**H12** 3-Hydroxy-2-Butanone

Supplementary Figure 4| BiOLOG MicroArray raw data. Duplicate experiments were performed for each plate, while time points were collected at 0, 3, 6, 12, 24, 36 and 24 hrs. The mean and standard deviation are presented. A major phenotypic change between strains is defined as a difference in the rate of respiration by <50% after 24 h growth (late log phase), on the same carbon source.

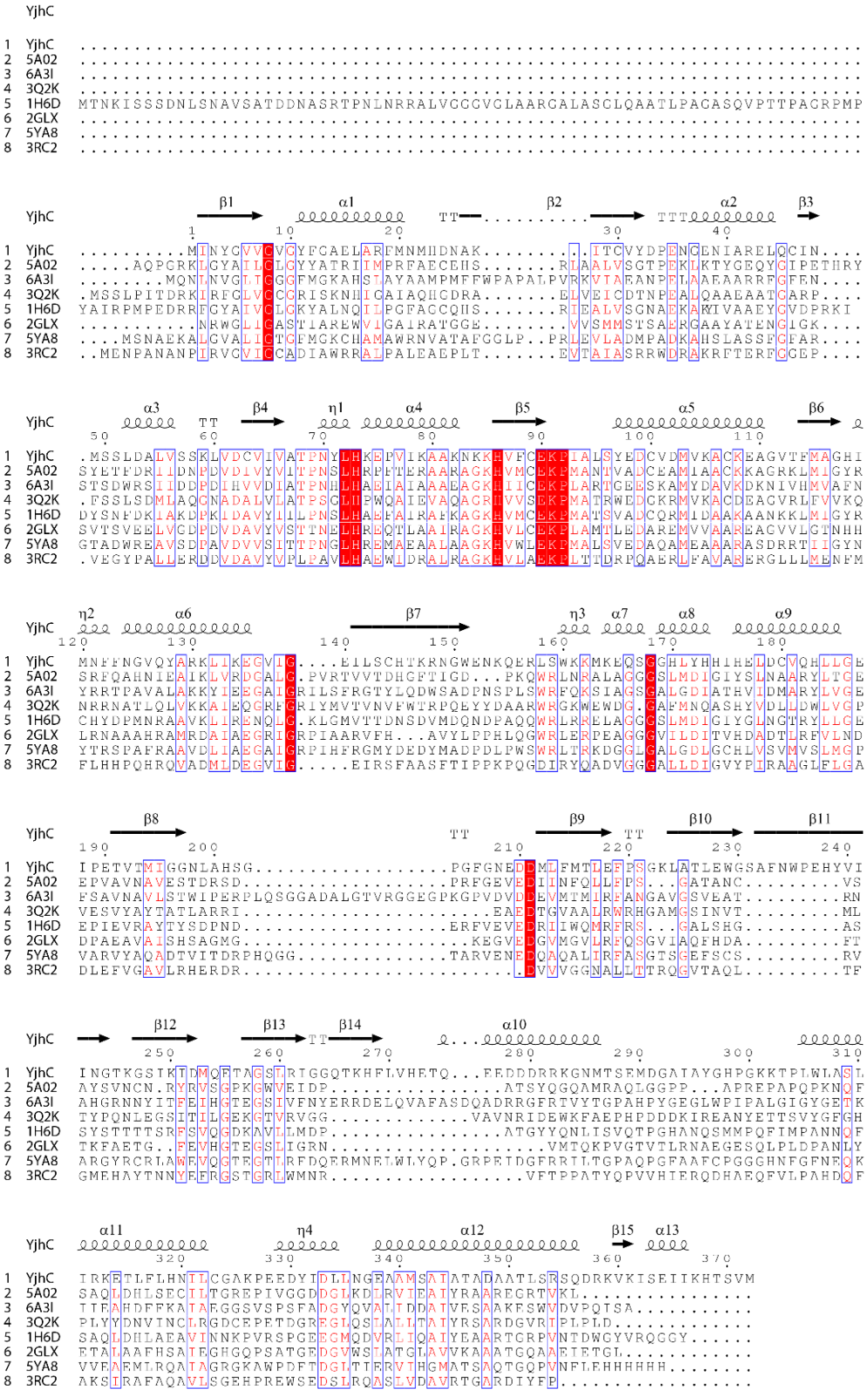
Supplementary Figure 5 | LigPlot highlighting the interactions that stabilise the cofactor NAD(H) in the substrate binding pocket. The hydrogen bonded network is depicted with dashed lines with the distance between partners labelled. Key water molecules are coloured in light blue, while hydrophobic interactions which stabilise NAD(H) are depicted as semi circles with dashes.



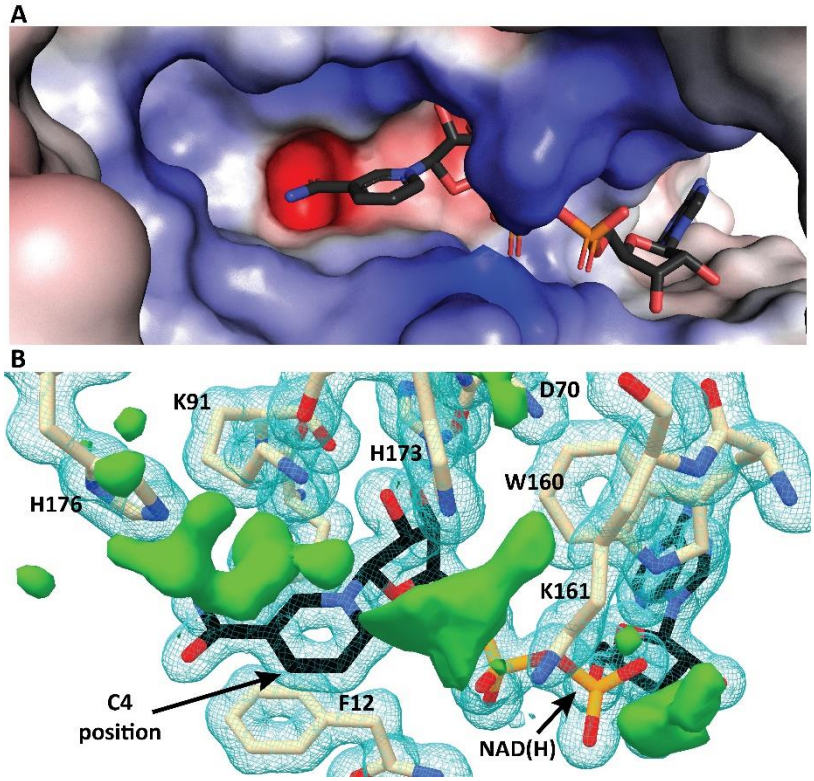
Supplementary Figure 6 | Molecular surface representations of the *E. coli* YjhC dimer interface. Each monomer is separated and rotated by 60° to expose features of their molecular surface in more detail. **A)** Two interacting monomers are shown juxtaposed with the monomer as a protein cartoon. Interface residues that form hydrophobic interactions are highlighted in red for both chains, while residues that form hydrogen bonding interactions or salt bridges are highlighted in blue and green from chain A and B, respectively. **B)** Two interacting monomers are coloured by their electrostatic potential. Figures produced in *PyMOL* (67).



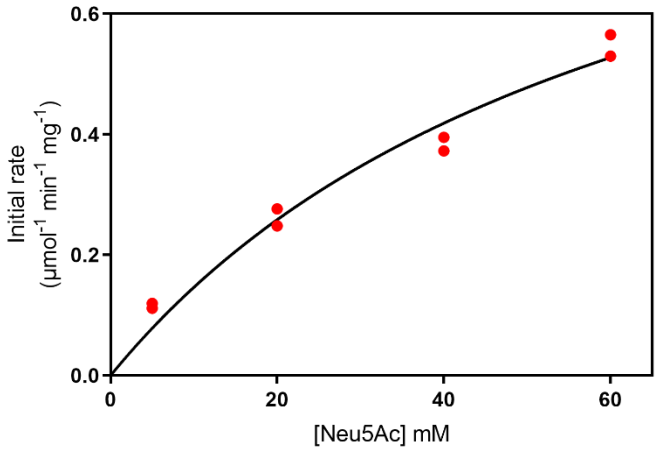
Supplementary Figure 7 | Multiple sequence alignment performed using *Clustal Omega* (23) and generated using *ESPript 3.0* (24). The top seven search hits from the *DALI* server (Table 3) were used in the sequence alignment, identified by their PDB I.D. Red background highlights identical residues, red coloured font indicates similar residues. Secondary structure elements depicted above the alignment were generated from the crystal structure of YjhC (**Figure 3**). The dots above the aligned sequences are the amino acid numbers respective to YjhC.



Supplementary Figure 8 | Binding cavity of YjhC. A) The substrate binding cavity mapped by electrostatic potential depicts primarily, a positively charged pocket rendering using the Adaptive Poisson-Boltzmann Solver (ABPS) plugin. Colored from red (negative) to blue (positive). **B)** Electron density map highlighting the unmodelled, ambiguous mF_o-DF_c difference density (3.0 σ, green). The $2F_o$-DF_c electron density map (1.0 σ, blue mesh) is shown around conserved residues within the binding cavity and the bound cofactor, NAD(H) (C4 position is labelled). Figures were produced using *PyMOL* (67) and *UCSF Chimera* (44), respectively.



Supplementary Figure 9 | Neu5Ac kinetic assay. Using a spectrophotometric NAD^+ assay, YjhC activity in the presence of Neu5Ac was observed by measuring the change in absorbance at 340 nm. This gave a small rate for Neu5Ac across a mM concentration range and an apparent K_m of 65.1 mM (R^2 = 0.9637). The experiment was performed in duplicate, where initial rates (red spheres) were analyzed and fitted with the Michaelis-Menten model (black line) using *GraphPad Prism 8.2.0* (GraphPad Software, La Jolla California USA).



Supplementary Table 1 | Differential scanning fluorimetry of YjhC with potential carbon sources. Thermal shift assays were repeated in triplicate for each carbon source alone and in the presence of NAD(P). The average melting temperature (T_m) is presented for each condition along with the standard deviation. Significant increases in melting temperature are highlighted in bold; these are presented in Figure 2.

Molecule	T_m (°C)			
	-Cofactor	+ NAD^+	+NADH	+$NADP^+$
YjhC	46.51 ± 0.10	**52.36 ± 0.21**	**53.45 ± 0.10**	46.39 ± 0.18
N-Acetylneuraminic acid	46.68 ± 0.25	**55.22 ± 0.10**	**56.23 ± 0.02**	X †
N-Acetylmannosamine	46.53 ± 0.20	52.54 ± 0.27	53.71 ± 0.14	X
N-Acetylglucosamine	46.54 ± 0.18	52.60 ± 0.18	53.44 ± 0.05	X
N-Acetylgalactosamine	46.42 ± 0.38	52.61 ± 0.18	53.61 ± 0.08	X
Sialic acid lactone	46.50 ± 0.15	**52.19 ± 0.29**	**54.41 ± 0.10**	X
Arabinose	46.60 ± 0.10	52.48 ± 0.21	53.50 ± 0.16	X
Trehalose	46.53 ± 0.20	52.84 ± 0.28	53.62 ± 0.07	X
Mannose	46.54 ± 0.32	52.60 ± 0.18	53.62 ± 0.09	X
Mannitol	46.54 ± 0.18	52.60 ± 0.32	53.66 ± 0.02	X
Glucose	46.58 ± 0.13	52.66 ± 0.10	53.44 ± 0.11	X
Maltose	46.60 ± 0.10	52.60 ± 0.18	53.55 ± 0.02	X
Galactonic acid-γ-lactone	45.99 ± 0.19	52.53 ± 0.47	53.49 ± 0.09	X

† Not tested, as +NADPH provided no increase in T_m

Supplementary Table 2 | Ensemble refinement statistics [a].

Refinement statistics	
Final single-structure model	
R_{work} (%)	14.98
R_{free} (%)	16.87
Geometric RMSD	
Bond (Å)	0.009
Angle (°)	1.310
Ensemble refinement model	
R_{work} (%)	13.08 (-1.90)
R_{free} (%)	15.93 (-0.94)
τ_x (ps)	1.0
Number of ensemble structures	100
Geometric RMSD (centroid distribution) †	
Bond (Å)	0.009
Angle (°)	0.984
Geometric RMSD (per model distribution) ‡	
Bond (Å)	0.016
Angle (°)	1.758

[a] Values in parentheses in the change in *R* factors between the ensemble and single model

† Calculated for the whole ensemble where restraints represent an oscillation around the ideal value

‡ Calculated for each model separately and then averaged across the whole ensemble

RMSD = root mean square deviation

Supplementary Table 3 | Summary of *in silico* docking experiments using *SwissDock.*

Molecule	Estimated Δ*G* *(kcal mol^{-1})*	Full Fitness *(kcal mol^{-1})*	Number of hydrogen bonds
N-Acetylneuraminic acid (β-anomer)	-5.98	-1906.47	2
N-Acetylneuraminic acid (Open chain)	-7.08	-1929.16	3
Sialic acid 1,7 lactone	-5.41	-1896.51	2
9-O-Acetyl-sialic acid	-9.4	-1902.13	4

2.4 Chapter summary

This chapter developed a structural and functional understanding of *E. coli* YjhC to address the hypothesis that the functional role of this previously uncharacterised protein was to process the less common derivatives of sialic acid.

Phenotype MicroArrays and differential scanning fluorimetry experiments demonstrated that *E. coli* YjhC is sensitive to carbohydrate-based substrates and identified Neu5Ac and derivatives of sialic acid as promising leads. Solution studies using sedimentation velocity experiments and SAXS verified that YjhC adopts a dimeric architecture in solution.

The reported high-resolution structure of *E. coli* YjhC, solved in the complex with the cofactor NAD(H) is consistent with its role as an oxidoreductase/dehydrogenase from the Gfo/Idh/MocA family. Further ensemble refinement identified a flexible loop region that may play a key role during catalysis. A structural comparison suggested this flexible loop was a unique feature to YjhC among closely related structures. Using *in silico* docking, Neu5Ac and various sialic acid derivatives were screening in order to identify molecules that presented favourable binding modes. This reinforced Neu5Ac as a promising lead and further established the lactone derivative and the 9-O-acetylated derivative of sialic acid as potential substrates.

This study provided the first characterisation of YjhC, indicating an involvement in sialic acid catabolism, which was consistent with the hypothesis in the literature (15), and verified its role as an oxidoreductase/ dehydrogenase. Together, this knowledge expands our understanding of bacterial sialic acid catabolism and provides a foundation towards antimicrobial drug development for this enzyme.

2.5 Chapter References

61. Kalivoda, K. A., Steenbergen, S. M., Vimr, E. R., & Plumbridge, J. (2003). Regulation of sialic acid catabolism by the DNA binding protein NanR in *Escherichia coli*. *Journal of Bacteriology, 185*, 4806-4815.
62. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
63. Traving, C., & Schauer, R. (1998). Structure, function and metabolism of sialic acids. *Cellular and Molecular Life Sciences, 54*, 1330-1349.
64. Almagro-Moreno, S., & Boyd, E. F. (2009). Insights into the evolution of sialic acid catabolism among bacteria. *BioMed Central Evolutionary Biology, 9*, 118.
65. Plumbridge, J., & Vimr, E. (1999). Convergent pathways for utilisation of the amino sugars *N*-acetylglucosamine, *N*-acetylmannosamine, and *N*-acetylneuraminic acid by *Escherichia coli*. *Journal of Bacteriology, 181*, 47-54.
66. Severi, E., Hood, D. W., & Thomas, G. H. (2007). Sialic acid utilisation by bacterial pathogens. *Microbiology, 153*, 2817-2822.
67. Teplyakov, A., Obmolova, G., Toedt, J., Galperin, M. Y., & Gilliland, G. L. (2005). Crystal Structure of the Bacterial YhcH Protein Indicates a Role in Sialic Acid Catabolism. *Journal of Bacteriology, 187*, 5520-5527.

68. Howard, R. J., Reuter, G., Barnwell, J. W., & Schauer, R. (1986). Sialoglycoproteins and sialic acids of Plasmodium knowlesi schizont-infected erythrocytes and normal rhesus monkey erythrocytes. *Parasitology, 92 (Pt 3)*, 527-543.
69. Muchmore, E. A., Diaz, S., & Varki, A. (1998). A structural difference between the cell surfaces of humans and the great apes. *American Journal of Physical Anthropology, 107*, 187-198.
70. Condemine, G., Berrier, C., Plumbridge, J., & Ghazi, A. (2005). Function and expression of an *N*-acetylneuraminic acid-inducible outer membrane channel in *Escherichia coli*. *Journal of Bacteriology, 187*, 1959-1965.
71. Wirth, C., Condemine, G., Boiteux, C., Berneche, S., Schirmer, T., & Peneff, C. M. (2009). NanC crystal structure, a model for outer-membrane channels of the acidic sugar-specific KdgM porin family. *Journal of Molecular Biology, 394*, 718-731.
72. Severi, E., Müller, A., Potts, J. R., Leech, A., Williamson, D., Wilson, K. S., & Thomas, G. H. (2008). Sialic acid mutarotation is catalyzed by the *Escherichia coli* β-propeller protein YjhT. *Journal of Biological Chemistry, 283*, 4841-4849.
73. Steenbergen, S. M., Jirik, J. L., & Vimr, E. R. (2009). YjhS (NanS) is required for *Escherichia coli* to grow on 9-O-acetylated N-acetylneuraminic acid. *Journal of Bacteriology, 191*, 7134-7139.
74. Rangarajan, E. S., Ruane, K. M., Proteau, A., Schrag, J. D., Valladares, R., Gonzalez, C. F., Gilbert, M., Yakunin, A. F., & Cygler, M. (2011). Structural and enzymatic characterization of NanS (YjhS), a 9-O-Acetyl N-acetylneuraminic acid esterase from *Escherichia coli* O157:H7. *Protein Science, 20*, 1208-1219.
75. Vimr, E. R. (2013). Unified theory of bacterial sialometabolism: how and why bacteria metabolise host sialic acids. *International Scholarly Research Notices Microbiology, 2013*, 816713.
76. Chang, D. E., Smalley, D. J., Tucker, D. L., Leatham, M. P., Norris, W. E., Stevenson, S. J., Anderson, A. B., Grissom, J. E., Laux, D. C., Cohen, P. S., & Conway, T. (2004). Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 7427-7432.
77. Almagro-Moreno, S., & Boyd, E. F. (2009). Sialic acid catabolism confers a competitive advantage to pathogenic *Vibrio cholerae* in the mouse intestine. *Infection and Immunity, 77*, 3807-3816.
78. Jeong, H. G., Oh, M. H., Kim, B. S., Lee, M. Y., Han, H. J., & Choi, S. H. (2009). The capability of catabolic utilisation of *N*-acetylneuraminic acid, a sialic acid, is essential for *Vibrio vulnificus* pathogenesis. *Infection and Immunity, 77*, 3209-3217.
79. Pezzicoli, A., Ruggiero, P., Amerighi, F., Telford, J. L., & Soriani, M. (2012). Exogenous sialic acid transport contributes to group B streptococcus infection of mucosal surfaces. *Journal of Infectious Diseases, 206*, 924-931.
80. Taberman, H., Parkkinen, T., & Rouvinen, J. (2016). Structural and functional features of the NAD(P) dependent Gfo/Idh/MocA protein family oxidoreductases. *Protein Science, 25*, 778-786.
81. Taberman, H., Andberg, M., Koivula, A., Hakulinen, N., Penttila, M., Rouvinen, J., & Parkkinen, T. (2015). Structure and function of *Caulobacter crescentus* aldose-aldose oxidoreductase. *Biochemical Journal, 472*, 297-307.
82. Nurizzo, D., Halbig, D., Sprenger, G. A., & Baker, E. N. (2001). Crystal structures of the precursor form of glucose-fructose oxidoreductase from *Zymomonas mobilis* and its complexes with bound ligands. *Biochemistry, 40*, 13857-13867.
83. Rowland, P., Basak, A. K., Gover, S., Levy, H. R., & Adams, M. J. (1994). The 3-dimensional structure of glucose-6-phosphate-dehydrogenase from *Leuconostoc mesenteroides* refined at 2.0-angstrom resolution. *Structure, 2*, 1073-1087.
84. Thoden, J. B., Sellick, C. A., Reece, R. J., & Holden, H. M. (2007). Understanding a transcriptional paradigm at the molecular level - The structure of yeast Gal80p. *Journal of Biological Chemistry, 282*, 1534-1538.

Chapter Three

On the structure and function of the proteins encoded by the *yjhBC* operon – Part Two

3.1 Introduction

3.1.1 The Major Facilitator Superfamily of permeases

The major facilitator superfamily (MFS) is a diverse group of permeases that permit the uptake of small solutes, such as sugars, amino acids or drugs across the cytoplasmic membrane within archaea, bacterial and eukaryotic species (1; 2). Based on their mode of uptake, members of the MFS are further classified as secondary symporters, secondary antiporters or uniporters (3). Secondary symporters facilitate transport by coupling the movement of a proton/ion down an electrochemical gradient to entropically drive the small solute against an electrochemical and/ or concentration gradient. Secondary antiporters facilitate a similar coupled process, although pump the different solutes in opposing directions. In contrast, uniporters simply utilise the electrochemical gradient of the small solute to entropically facilitate transport (4; 5). These transport modes are illustrated in Figure 3.1.

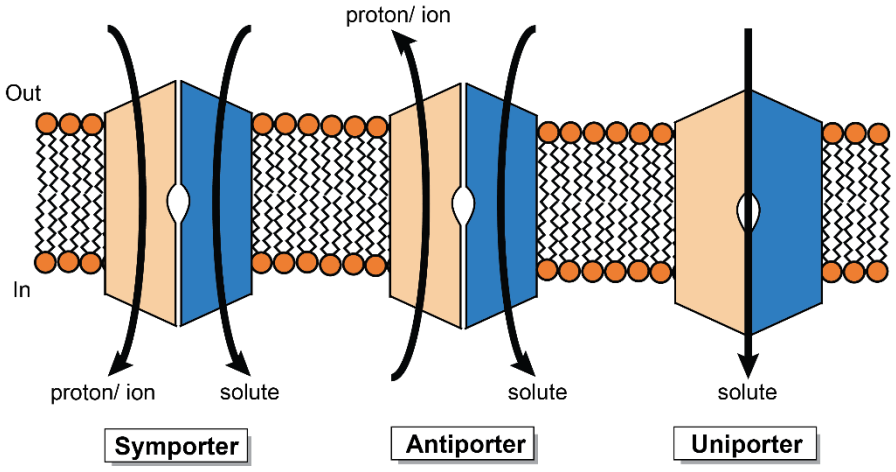


Figure 3.1 | Transport modes catalysed by members of the major facilitator superfamily. This model presents the secondary symporter, secondary antiporter and uniporter modes of uptake across the plasma membrane. The black arrow indicates the direction of transport for the solute and co-solute (proton/ ion).

MFS permeases are characterised by a topology of 12 transmembrane α-helices that are shared between a structurally similar N- and C-terminal domain (6; 7). Exceptions to this trademark feature include a member with six transmembrane α-helices, which likely functions as a dimer (1), and a member with 24 transmembrane α-helices that is likely the result of a gene duplication and fusion (1; 8). In addition, other members of the MFS have been identified to contain 14 transmembrane α-helices. An example is the sialic acid-proton symporter NanT, where the additional transmembrane helices, located centrally are hypothesised to form an amphipathic helix and play a role in the substrate specificity of the transporter (9).

Despite a difference in transport modes (Figure 3.1), MFS members are proposed to function *via* a common mechanism, which is described as the 'rocker-switch' model. In this model, transporters are known to alternate between an inward- and outward-facing conformation to move molecules across a membrane, where their respect substrate-binding site is only accessible to either the cytoplasmic or periplasmic side of the membrane (Figure 3.2) (10; 11). The substrate-binding site or translocation pathway is located at the domain interface between the N- and C- α-helical bundles. This alternating access between two conformations in order to facilitate transport has been well characterised in the *Escherichia coli* glycerol-3-phosphate inorganic phosphate antiporter (11) and the *E. coli* lactose permease sugar proton symporter (10).

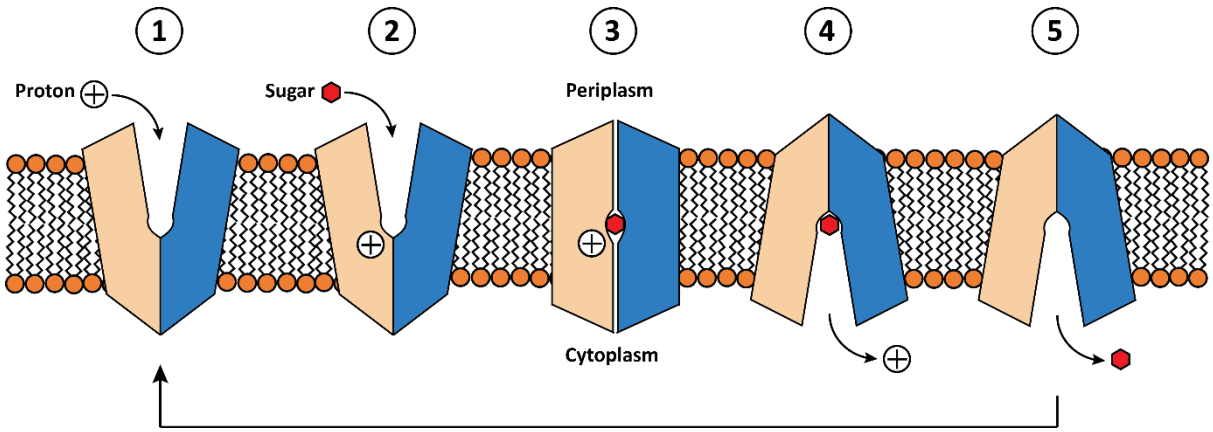


Figure 3.2 | The 'rocker-switch' model of transport within the major facilitator superfamily. In this mechanism the N- and C-terminal domains (coloured in beige and blue, respectively) rock back and forth between an inward- and outward-facing open conformation, where the substrate-binding site acts as the pivot point, accessible at either the cytoplasmic or periplasmic side of the membrane at a given time. Initially, in an outward-facing open conformation a proton binds from the periplasmic side of the membrane (Position 1). Following the binding of a proton, the substrate binds in a central cavity (Position 2). The binding of the substrate triggers a conformational change, initially through an intermediate occluded state (Position 3), through to an inward-facing open conformation where the proton can exit into the cytoplasm (Position 4). Finally, the substrate is translocated and released into the cytoplasm, completing the cycle (Position 5). [Figure adapted from (12)].

3.1.2 The putative permease YjhB

Encoded by the *yjhBC* operon, YjhB is placed in the MFS and is hypothesised to function as a permease, playing a role in sialic acid catabolism (13). This hypothesis follows the identification that the *yjhBC* operon contains three direct repeats of the nucleotide sequence GGTATA, which is specifically recognised by NanR, the transcriptional regulator of sialic acid catabolism in *E. coli* (14). Furthermore, YjhB is comparable to the sialic acid-proton symporter NanT (13). This suggests that YjhB is not the main transporter for sialic acid, as studies have shown that NanT knockouts prevent growth on sialic acid (15) and therefore supports the hypothesis that YjhB functions to import less common derivatives of sialic acid (13). However, there is no structural or functional data to verify this hypothesis.

3.2 Chapter overview

To expand on the characterisation of the *yjhBC* operon (Chapter Two), this chapter investigates the putative permease YjhB. The results presented here show the overexpression trials conducted for YjhB and conclude an optimised protocol for expression. Furthermore, *in silico* comparison against sequence homologs was used to identify potential amino acid residues important for function. Following reconstruction of a *de novo* 3D model, these amino acid residues were mapped to the structure. Together, Chapter Two and Three provide invaluable insight into this uncharacterised operon.

3.3 Results and Discussion

3.3.1 Cloning and overexpression trials of *Escherichia coli* YjhB

The structural and functional characterisation of proteins is increasingly dependent on purifying high amounts of pure and stable protein, where some problems are related to gene expression (16; 17). The overexpression of membrane proteins is often toxic to the host, caused by saturating the protein secretion machinery and membrane protein biogenesis (18; 19). Thus, efficient methods aimed at controlling the expression of membrane proteins and preventing toxicity are used to obtain sufficient quantities of pure protein required for their structural and biophysical characterisation. These methods utilise expression systems with a tightly regulated promoter, avoiding systems driven exclusively by the bacteriophage T7-ribonucleic acid (RNA) polymerase, for example, the *E. coli* BL21(DE3) strain, which can result in leaky expression (20).

To optimise expression of the *yjhB* gene, the *E. coli* Lemo21(DE3) strain was utilised. This is a T7-RNA polymerase-based platform for the overexpression of challenging targets, such as membrane proteins, proteins with solubility issues or toxic proteins (21). This strain was developed from the *E. coli* BL21(DE3) strain, one of the most extensively used platforms to recombinantly overexpress protein in *E. coli* (18). Governed by the isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible *lac*UV5 promoter, protein overexpression is driven by T7-RNA polymerase to transcribe messenger RNA approximately five times faster than the *E. coli* RNA polymerase (18). This feature, although attractive to routine protein expression is not well suited to challenging targets, such as membrane proteins. To address this obstacle the derivative strain of *E. coli* BL21(DE3), known as Lemo21(DE3), was developed to control the rate of expression by using T7 lysozyme, a natural inhibitor of T7-RNA polymerase. Encoded on a separate plasmid identified as pLemo, the expression of T7 lysozyme is governed by the titratable rhamnose (*rha*BAD) promoter (Figure 3.3) (22). The rationale behind this titratable system is that varying the amount of L-rhamnose controls the amount of protein expressed and reduces the risk of saturating the protein secretion machinery and membrane protein biogenesis (18; 19).

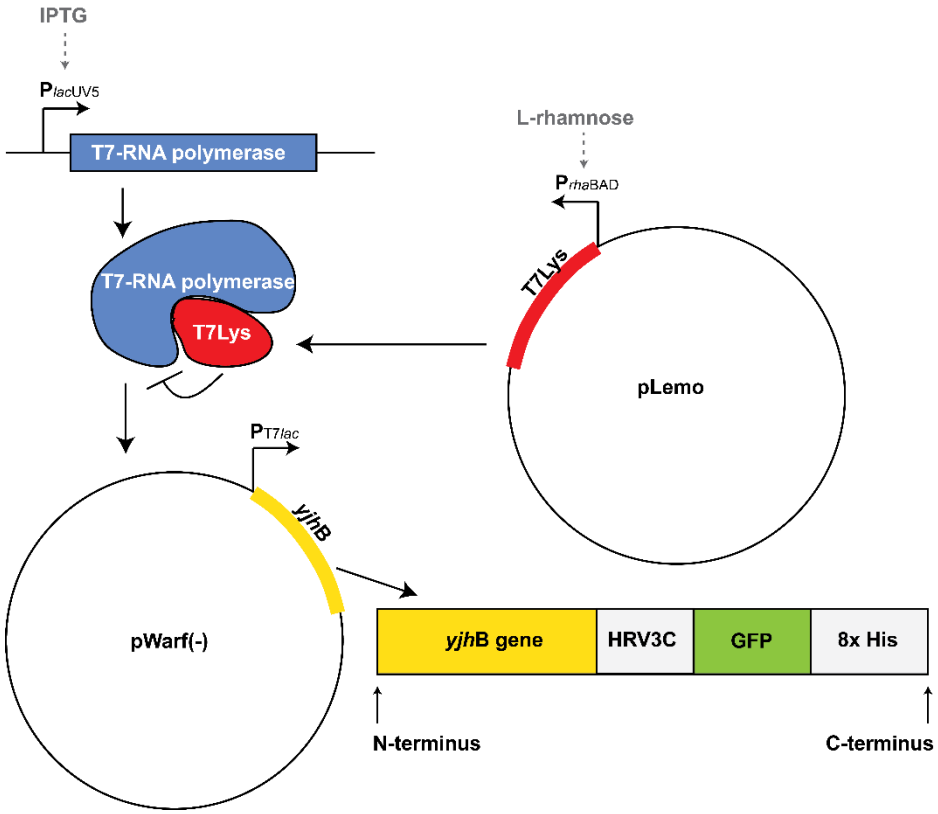


Figure 3.3| The *E. coli* Lemo21(DE3) strain and pWarf(-) expression vector allow for tunable expression of *yjhB*. The presence of IPTG governs the expression of T7-RNA polymerase under the *lac*UV5 promoter (P_{lacUV5}). The activity of T7-RNA polymerase is controlled by its natural inhibitor, T7-lysozyme (T7Lys). Present on the pLemo plasmid, the expression of T7Lys is governed by the titratable *rha*BAD promoter (P_{rhaBAD}) using L-rhamnose. Ultimately this controls the overexpression of *yjhB* located on the pWarf(-) expression vector, governed by the T7 lac promoter (P_{T7lac}). The pWarf(-) expression vector is shown [Figure adapted from (18)].

To monitor the overexpression of membrane proteins, the *E. coli* Lemo21(DE3) strain was used in combination with the commercially available low-copy pWarf(-) expression vector. The vector carries a C-terminal human rhinovirus 3C protease (HRV3C) cleavage site, a C-terminal green fluorescent protein (GFP) and a C-terminal 8x histidine affinity tag (Figure 3.3) (23). The GFP fusion tag permits the levels of overexpressed membrane protein to be monitored by using whole-cell fluorescence, a high-throughput method to decipher the optimal expression conditions of the target protein compared to traditional and more tedious approaches such as immunoblotting. Put simply, the higher the fluorescence, the more stable the protein expressed. If no fluorescence is observed, the target protein is not correctly folded and expression conditions need to be revised. As a consequence, the *E. coli* Lemo21(DE3) strain and pWarf(-) expression vector allow for the tunable expression of *yjhB*. Utilisation of pWarf(-) is only possible if the C-terminus of the protein is intracellular, as GFP is only fluorescent in the cytoplasm. If the C-terminus is extracellular or positioned in the periplasm, the expression vector pWarf(+) can be utilised, which repositions the C-terminus to the cytoplasm using a Glycophorin A protein fusion (23). The sub-cloning of *E. coli yjhB* into the pWarf(-) expression vector and subsequent transformation into *E. coli* Lemo21(DE3) is described in Chapter Eight, Section 8.6.1.1.

Using a high-throughput screen developed by Drew *et al.* 2006, YjhB was subjected to a total of 48 different expression variables in order to find the optimal overexpression conditions for the protein (24). These small-scale expression trials were carried out in terrific broth (TB) media, where a range of L-rhamnose concentrations (0-2000 µM), induction temperatures (18 and 26 °C) and IPTG concentrations (0.1, 0.4, 0.7 and 1 mM) were explored. All conditions had a total induction time of 16 h. As a negative control, empty pWarf(-) was included in the expression screen as the presence of a target gene is required for GFP to be expressed. The expression of each variable was analysed by whole-cell fluorescence (Chapter Eight, Section 8.6.1.2), where the best range of overexpression conditions were visualised by in-gel fluorescence using gel electrophoresis (Chapter Eight, Section 8.6.1.3). The whole-cell fluorescence measurements for each expression screen are presented in Figure 3.4A, where the highest whole-cell fluorescence was observed with 250 µM L-rhamnose and 0.1 mM IPTG, at 26 °C. In addition, the second highest whole-cell fluorescence was observed with 250 µM L-rhamnose and 1.0 mM IPTG, at 26 °C. Both of these overexpression conditions displayed a relative fluorescence unit (RFU) above 15, which represents the minimum level of expression required to isolate a suitable yeild of protein (at least 1 mg L^{-1}) for structural and biophysical studies (23). When visualised using in-gel fluorescence, YjhB appeared at approximately 40 kDa (Figure 3.4B), although the predicted weight of the YjhB-GFP fusion protein is 73.4 kDa. This anomalous gel migration is commonly observed in SDS-PAGE when running membrane proteins and is believed to be attributable to the binding of detergent molecules to the transmembrane regions (25). Of the conditions analysed using in-gel fluorescence, the

most intense band correlates with 250 µM L-rhamnose and 0.1 mM IPTG, at 26 °C (Figure 3.4B), which was the best optimisation condition as determined by whole-cell fluorescence. The same overexpression condition is shown side-by-side to an uninduced sample (resuspended cell pellets), in order to illustrate the difference between the inherent fluorescence of *E. coli* and the fluorescence observed with overexpression of the YjhB-GFP fusion (Figure 3.4C). Taken together, this condition is suitable for protein overexpression at a larger scale, due to having an RFU above 15.

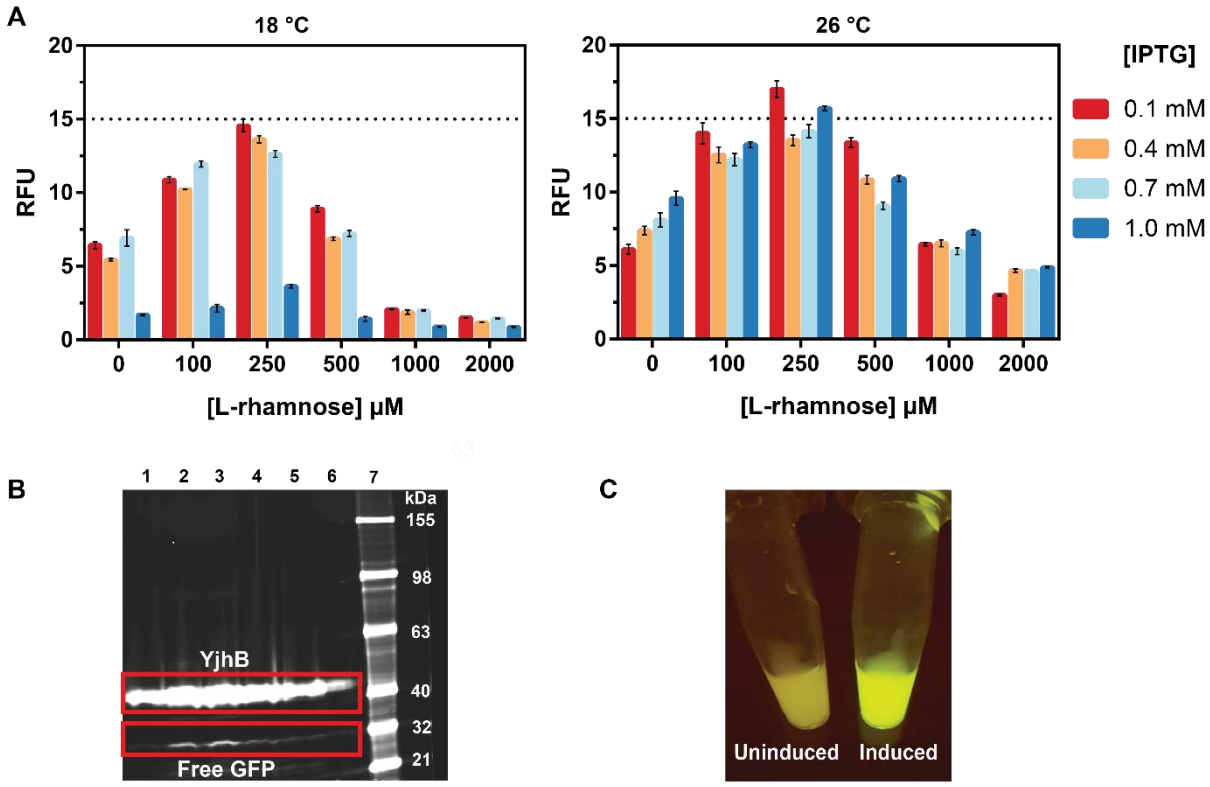


Figure 3.4 | Whole cell fluorescence and in-gel fluorescence of YjhB in different expression conditions. A) Whole cell fluorescence for two induction temperatures (18 and 26 °C) reported in relative fluorescence units (RFU). YjhB was subjected to four different IPTG concentrations (0.1, 0.4, 0.7 and 1.0 mM), across a range of L-rhamnose concentrations (0, 100, 250, 500, 1000 and 2000 µM). Each IPTG concentration is coloured and grouped with each respective L-rhamnose concentration. The standard deviation was calculated from the duplicate measurements and shown as error bars. The horizontal dashed line highlights 15 RFU representing the minimum level of expression required to isolate a suitable amount of protein (at least 1 mg L^{-1}). **B)** In-gel fluorescence of YjhB from the best overexpression range at 26 °C using an IPTG concentration of 0.1 mM, as determined by whole-cell fluorescence. Lane 1 – 0 µM L-rhamnose; Lane 2 – 100 µM L-rhamnose; Lane 3 – 250 µM L-rhamnose; Lane 4 – 500 µM L-rhamnose; Lane 5 – 1000 µM L-rhamnose; Lane 6 – 2000 µM L-rhamnose; and Lane 7 – BenchMark Fluorescent Protein Standard (kDa). Although a tiny amount of free GFP is observed in the gel, the YjhB-GFP fusion protein is detected at levels that are sufficient for structural and biophysical studies (conditions that are above 15 RFU) **C)** The difference between the inherent fluorescence of *E. coli* (uninduced) and the fluorescence of the best overexpression condition, 250 µM L-rhamnose and 0.1 mM IPTG, at 26 °C.

Although not addressed in this thesis, the next step towards characterising YjhB would be to optimise a protein purification protocol. To initiate this process, the membrane fraction must be harvested by ultracentrifugation and subsequently solubilised using detergents. Identifying a suitable detergent that will mimic the native membrane environment for the protein is a major bottleneck in obtaining stable and monodisperse protein (23). Traditionally, this process was lab intensive and required microgram to milligram quantities of purified protein (23). However, fluorescence-detection size-exclusion chromatography (FSEC) has recently emerged as an efficient strategy to identify the optimal detergent that renders the protein stable in solution by exploiting the GFP-fusion tag and using only nanogram quantities (26). As a result, crude membrane extract can be solubilised in a range of detergents and subjected to FSEC, equilibrated in the respective detergent-containing buffer. At this point, using fluorescence spectroscopy, both the stability and monodispersity of the GFP-fusion protein can be analysed for each respective detergent by monitoring the elution profile over time. Here, a single, symmetrical peak represents stable, monodisperse protein, while in contrast multiple or largely asymmetric peaks are reflective of an unstable and polydisperse protein (23; 26). The ideal detergent will present a stable elution profile over a 24 to 48 h period.

3.3.2 *In silico* comparison and structural modelling of YjhB

Since YjhB is yet to be purified, bioinformatics were used to provide structural and functional insight for this uncharacterised permease. To achieve this, various tools were employed to search for appropriate templates, construct a *de novo* 3D model, predict how this model may embed within the lipid bilayer and identify potential residues that are important to its biological function.

Initially, a sequence-based homology search was conducted using the BLASTp web server to verify the annotation of YjhB in the MFS of permeases. Of the homologs producing significant alignment with YjhB, the sialic acid-proton symporter NanT was most closely related. Interestingly, *E. coli* YjhB presented the highest sequence identity to NanT from *Yersinia* species (37.1%), followed closely by *Salmonella enterica* sub-species (35.2%), rather than *E. coli* NanT, which was third highest (35.0%). The top five results of this homology search from the UniProt database are presented in a multiple sequence alignment (Figure 3.5). Collectively, these all represent NanT from various bacterial species, while the other homologs with a sequence identity less than 30% represent an assortment of putative sugar-, amino acid- and metabolic-proton symporters (27-30). This reinforces the annotation of a permease made previously for YjhB (13), and proposes that perhaps YjhB is really a NanT.

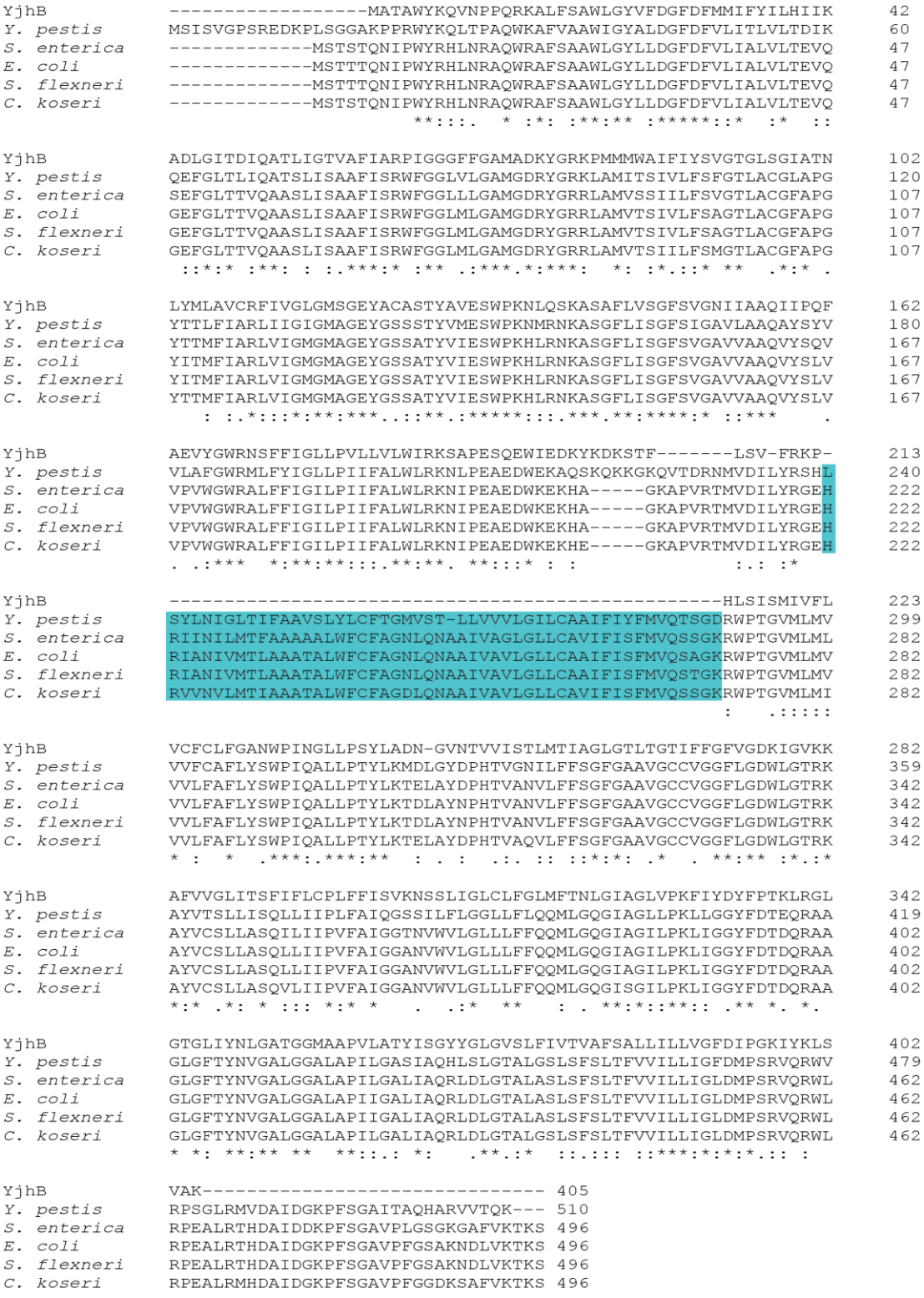


Figure 3.5 | Amino acid sequence alignment of YjhB compared to NanT homologs from five different species of bacteria. Species include *Yersinia pestis*, *Salmonella enterica*, *E. coli*, *Shigella flexneri* and *Citrobacter koseri*—all gram-negative bacteria. Strictly conserved residues are highlighted (*), conservation between residues with strongly similar properties are shown (:) and conservation between residues with weakly similar properties are shown (.). Residues implicated in the two additional and unique transmembrane helices of NanT are highlighted with a blue box.

The multiple sequence alignment of these transporters identified numerous regions of sequence conservation, including the DXXXGRR/K and E------R/Q motifs, which have been determined to play a crucial role in the well-characterised *E. coli* D-xylose sugar porter, XylE (31). One obvious difference is the extra 49 residues located halfway through the alignment for all NanT homologs. These residues likely comprise the two additional transmembrane helices that are unique to NanT (9). To confirm this assumption, the membrane topology of YjhB was predicted using the TOPCONS server (topcons.cbr.su.se/). Given an amino acid sequence, five different prediction algorithms are employed to specify the orientation of the transmembrane helices. Collectively, their predictions are supplied to the TOPCONS Hidden Markov Model, which provids a consensus prediction, coupled with a reliability score for the membrane topology (32). Based on the amino acid sequence, *E. coli* YjhB was predicted to comprise a total of 12 transmembrane helices, where both the N- and C-termini are located on the cytoplasmic side of the membrane (Figure 3.6). This is a characteristic feature of the major facilitator superfamily (33) and essential prerequisite for the pWarf(-) expression vector.

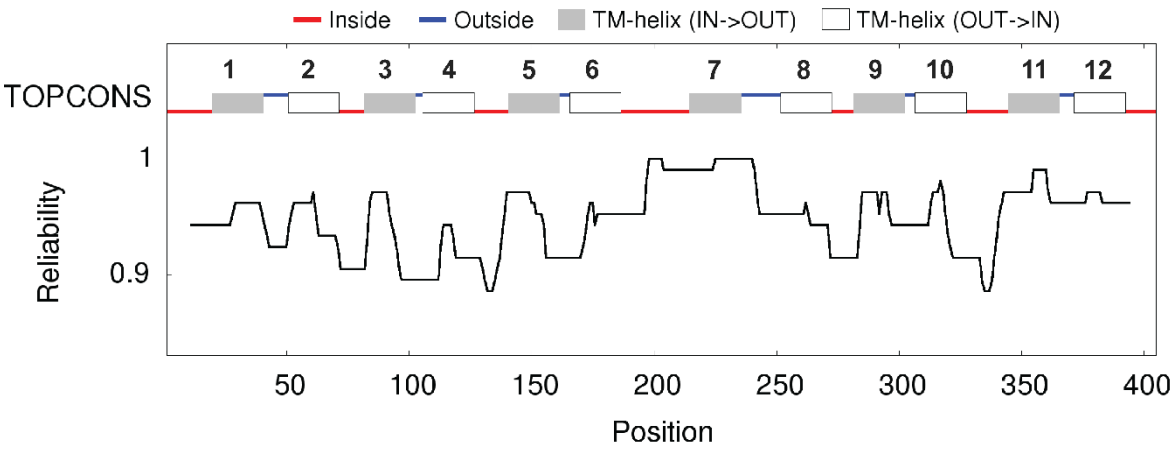


Figure 3.6 | The predicted membrane topology for YjhB. Using the TOPCONS server (32), the consensus prediction for the orientation of the transmembrane helices is presented. Transmembrane helices are numbered from 1 to 12, where their respective orientation is labelled. The reliability score is stated for each amino acid position in the sequence—a score of 1.0 reflects the highest quality. The N- and C- termini for YjhB are both predicted to be on the inside (or cytoplasmic side) of the plasma membrane.

To facilitate the characterisation of membrane proteins and identify functionally important features, the determination of the proteins 3D structure is essential, especially if the protein of interest is targeted for the development of therapeutics. Although membrane proteins constitute at least 20-30% of the total proteome for most organisms (34), they are severely under-represented in the PDB, currently at only 1-2% (2723 membrane proteins/ 150145 total PDB entries, March 2019). This paucity reflects the inherent challenges with protein overexpression, protein stability, method of reconstitution (i.e.

detergent, SMALPS, nanodiscs etc.) and simply obtaining well-ordered crystals for X-ray crystallography (18; 35). While the 'resolution revolution' in single-particle cryo-electron microscopy signals the beginning of a new area in structural biology (36; 37), this will still present challenges to membrane proteins, particularly with the difficulties in sample preparation and to a certain extent, image classification (38). As a result of these inherent challenges, computational approaches that predict the structure of membrane proteins still have a role in their characterisation (39; 40). Homology modelling is one computational approach to predict the structure of proteins through a template-based method, using closely-related atomic structures in the PDB (41). While this can be effective with soluble proteins, its success can be limited with membrane proteins due to their poor representation in the PDB, as mentioned above. An alternative is *de novo* protein structure prediction.

De novo protein structure prediction, as opposed to homology modelling constructs a 3D model from sequence or first principles, without the use of closely related structures or templates (42). To achieve this, two main approaches are commonly used; contact-assisted *ab initio* folding and fragment-based *ab initio* assembly. The contact-assisted approach aims to predict the spatial relationship between amino acid residues within a protein through a combined machine learning or evolutionary-coupled analysis (40; 43). Amino acid residues identified to have a spatial relationship, driven by a multiple sequence alignment, are used to generate a contact map and distance matrix. Combined with secondary structure information, these spatial relationships can then be used to construct a 3D model of the target protein (44). However, the accuracy of these models can be limiting when sequence homologs are lacking (45).

Alternatively, fragment-based assembly utilises small fragments of experimentally determined structures to construct an *ab initio* 3D model—a process driven by conformational sampling (46; 47). In comparison to soluble proteins, which display huge structural diversity, membrane proteins can be constrained to a conformational search space within the lipid membrane and present a simplified structural diversity of either α-helical bundles or β-barrels (48). Therefore, in the absence of structural homologs these small peptide fragments can be exploited where high local sequence identity is observed, in order to drive *ab initio* structure prediction with the assistance of simulated annealing and a scoring function to favour structures with protein-like features or folds (39; 49). This approach is fully implemented in the Rosetta software package (47; 50). However, the success of this strategy is dependent on a sequence identity of ≥30% and access to several structural homologs in the PDB (51). As the highest sequence identity was 23% for the structural homolog, D-xylose transporter XyIE from *E. coli* (PDB ID - 4GBY) (31), the contact-assisted *de novo* strategy was chosen over the fragment-based, comparative modelling strategy.

To construct a contact-assisted *de novo* 3D model for YjhB, the PredMP server was employed. This is a new state-of-the-art prediction method, which utilises a Deep Transfer Learning method, along with the latest features of the RaptorX web portal for protein structure and function prediction (52). Its workflow follows three modules: 1) prediction of secondary structural elements from the target amino acid sequence using the RaptorX-Property server, including any disordered regions (53); 2) the spatial relationship between residues are predicted through Deep Transfer Learning generating a contact map and distance matrix. This information is then combined with the predicted secondary structure to construct a *de novo* 3D model using the RaptorX-Contact server (40); and 3) The *de novo* 3D model is embedded into a lipid membrane, guided by a prediction of the membrane topology. A summary of this workflow is presented in Figure 3.7.

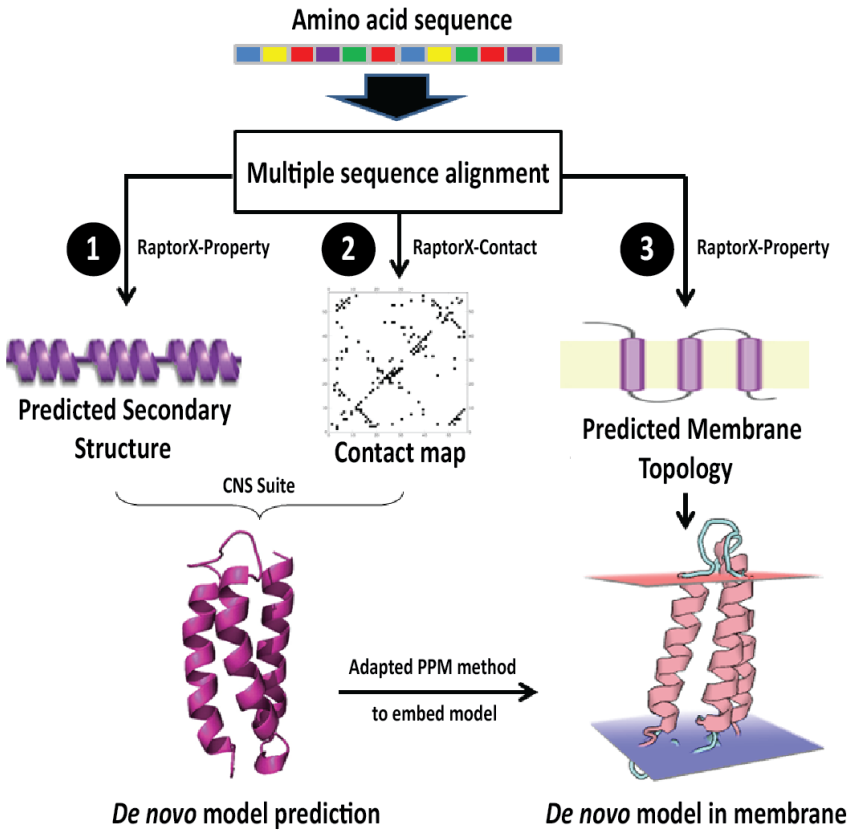


Figure 3.7 | *De novo* structure prediction workflow for PredMP. To construct a *de novo* 3D model the PredMP server initially requires the input of an amino acid sequence, which is used to construct a multiple sequence alignment with homologs. This sequence alignment is applied to a workflow of three modules: **1)** prediction of secondary structural elements; **2)** predict the contact map and combine this with the predicted secondary structure to construct a *de novo* 3D model using the 'Crystallography & NMR System' (CNS) Suite; and **3)** predict the membrane topology and insert the model into a lipid membrane using an adapted 'Positioning of Proteins in Membranes' (PPM) method. [Figure adapted from (52)].

The Deep Transfer Learning method significantly improves the quality of the *de novo* 3D model *via* increased accuracy when predicting the spatial relationship between residues (contact map), which is irrespective on the amount of sequence homologs available (45). To drive *de novo* 3D model construction, the RaptorX-Contact server (40) utilises the 'Crystallography & NMR System Suite'; a highly-flexible platform allowing structure-determination algorithms to easily be adapted for different systems such as PredMP (54). Consequently, this permits the distance matrix obtained from the contact map and the various geometric constraints (distance, angle and bonding) determined from the predicted secondary structure to be implemented in to the 'Crystallography & NMR System' Suite and guide *de novo* structure prediction (52). Finally, to embed the *de novo* 3D model into the lipid membrane, PredMP employs an adaption of the 'Positioning of Proteins in Membranes' method (55) to optimise the membrane burial potential and positioning of the transmembrane regions.

The output from the PredMP server generates a total of five full-length *de novo* 3D models, ranked according to their energy function from the 'Crystallography & NMR System' Suite (40). The estimated accuracy of the constructed models is defined by a confidence score. This confidence score is a measure of how much sequence information was available for the given protein to accurately construct a final 3D model. Given the term *Meff*, a protein with a score of 1.5 or less is marked with low confidence, a protein with a score of 3.5 or more is marked with high confidence, while a protein with a score in between is marked with medium confidence (45). Therefore, a low score reflects a protein with a poor representation of sequence homologs in which a fragment-based, comparative modelling strategy may offer a more accurate method of *de novo* structure prediction. The top *de novo* 3D model output from PredMP for YjhB had a *Meff* value of 3.125, which reflects medium confidence. However, it is largely skewed towards a level of high confidence, suggesting there is an acceptable number of sequence homologs for YjhB and the model can be trusted.

The top *de novo* 3D model for YjhB, presented in Figure 3.8A, displays the typical architecture for a member of the MFS, containing 12 transmembrane helices (supporting the TOPCONS prediction, Figure 3.6), which are shared equally in an N- and C-terminal domain, connected through a long cytoplasmic linker. Although this linker (along with the N- and C-terminal regions) was predicted to be partially disordered, it does contain a small cytoplasmic helix (C1) that perhaps may play a role in a gating mechanism during translocation of the substrate. Mapping the electrostatic potential to the surface of the *de novo* model, using the Adaptive Poisson-Boltzmann Solver (56), revealed a distribution of negative (red) charge on both the periplasmic and cytoplasmic sides of the membrane. This feature may facilitate the attraction of an amino sugar or proton towards its respective translocation pore in the protein. Characteristically, the substrate translocation pore in members of the MFS is located at the interface formed between the N- and C-terminal domains (33) (Figure 3.8A). As the contact-assisted *ab*

initio folding strategy does not allow conformational sampling (48), the *de novo* 3D model of YjhC, perhaps more closely resembles an occluded or intermediate state. Consequently, the surface cutaway representation (membrane view) did not expose any solvent accessible regions (Figure 3.8B), however the maximum dimensions of the protein could still be approximated.

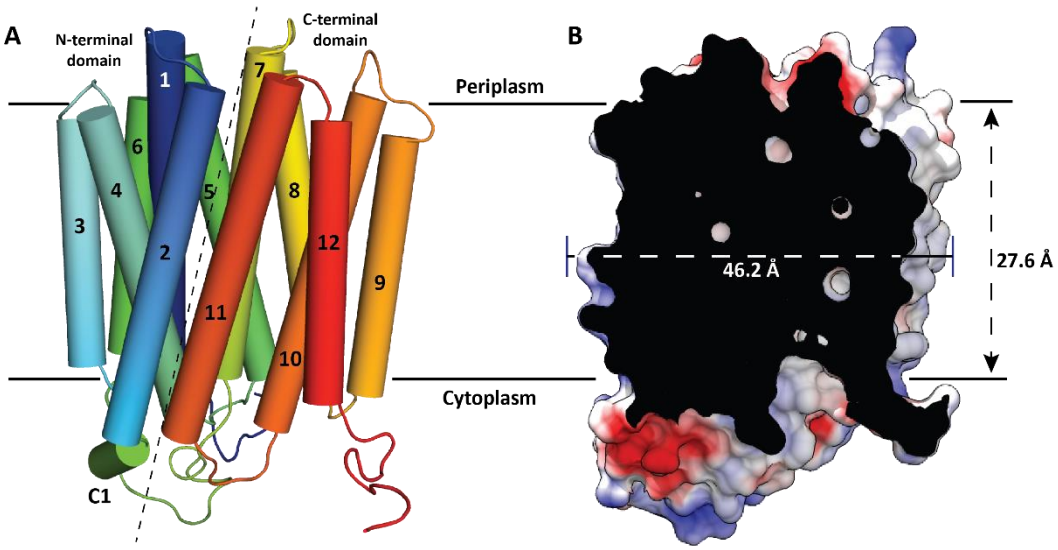


Figure 3.8 | *De novo* structure of YjhB. A) The overall structure of YjhB contains 12 transmembrane helices, characteristic of MFS members, these are depicted as cylinders, numbered and coloured by rainbow. The interface of the N- and C-terminal domains is highlighted (black dash). The cytoplasmic helix (C1) may play a role in a gating mechanism. **B)** A cutaway representation of the surface electrostatic potential is shown, which was calculated using the Adaptive Poisson-Boltzmann Solver (56).

Overall, the regions with the highest degree of sequence conservation are localised at the interface between the N- and C-terminal domains and therefore are likely to be involved in substrate translocation. To decipher regions of functional importance to YjhB, the sequence conservation of the top 50 homologs were mapped onto the *de novo* 3D model (Figure 3.9). These sequence homologs included NanT, which was used to generate the multiple sequence alignment in Figure 3.5, along with several other NanT homologs and an assortment of putative sugar-, amino acid- and metabolic-proton symporters, collectively with a sequence identity to YjhB of >20%. This visual representation of sequence conservation avoided having to interpret an extensive multiple sequence alignment and facilitated an ease of interpretation for consensus regions. Specifically, the motifs DGFD (transmembrane helix 1, residues 27-30), GxxxDK/RxGRK/R (between transmembrane helices 2 and 3, residues 70-80) and GWRxxF (transmembrane helix 6, residues 167-172) were highly conserved across all sequence homologs. The positively-charged residues of the GxxxDK/RxGRK/R motif are believed to interact with

the head-group of lipids, stabilising the protein in the membrane—a signature of sugar transporters (57; 58). The GWRxxF motif likely performs a similar role due to its location, interacting with lipids through charged and hydrophobic residues, while the DGFD motif is centrally located on transmembrane 1 where it may play a role in substrate binding. Interestingly, the cytoplasmic helix (C1) was identified to be highly variable, but without functional data and an experimentally solved structure it is difficult to determine if this feature is unique to YjhB.

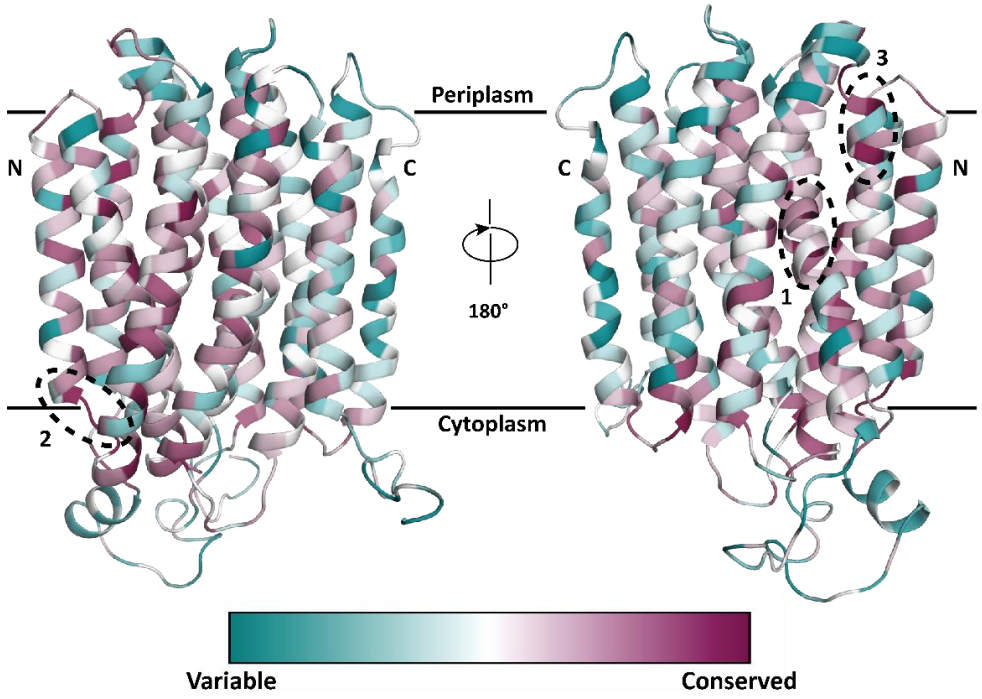


Figure 3.9 | Sequence conservation among homologs for YjhB. A multiple sequence alignment of the top 50 sequence homologs was generated using Clustal Omega (59). This was used to map the sequence conservation onto the *de novo* 3D model for YjhB using ConSurf (60). The conservation is colour-coded onto the structure as follows: highly conserved residues are maroon; residues that evolve at average rates are white; and residues that are highly variable are turquoise. Highlighted in black dashed circles are three highly conserved motifs: **1)** DGFD (transmembrane helix 1, residues 27-30); **2)** GxxxDK/RxGRK/R (between transmembrane helices 2 and 3, residues 70-80); and **3)** GWRxxF (transmembrane helix 6, residues 167-172).

As NanT is the closest sequence homolog to YjhB, it can be annotated as a sugar-proton/ion symporter that couples the movement of a proton/ion down an electrochemical gradient with the uphill movement of a sugar against an electrochemical and/ or concentration gradient in the same direction across the plasma membrane (33). To understand how this process may occur in YjhB, a structural similarity search was conducted. This method identified only a handful of structures, which was reflective on the small number of published structures available for major facilitator superfamily proteins. Despite these structural homologs having a low sequence identity and coverage (<24% identity and a query coverage

of 40-60%) they enabled functionally important residues to be identified and helped differentiate these residues between a potential substrate and proton/ion translocation pore.

The *E. coli* D-xylose transporter XyIE (PDB ID - 4GBY), a sugar-proton symporter within the MFS (31) was one of the closet structural homologs to YjhB and hence was considered the best structural candidate for both YjhB and NanT (7). The substrate of XylE, D-xylose, is bound between the N- and C-terminal domains within a central cavity and is coordinated by both aromatic and polar residues (31). The presence of these aromatic residues within the substrate binding cavity is a trademark feature of membrane transporter proteins (61-63). Within XylE, a total of four aromatic residues are situated near the substrate, which include Phe24, Tyr298, Trp392 and Trp416, located on transmembrane helices 1, 7, 10 and 11, respectively (31). In YjhB, these aromatic residues, although not explicitly conserved can be represented by Phe31, Phe226, Phe229, Trp233 and an additional Phe148 in the same vicinity (Figure 3.10). Furthermore, the identification of the polar residues Ser149, Asn152, Asn232, Asn236 and Thr264 surrounding these aromatics collectively would support the coordination of a sugar molecule at this location. On this basis, I propose that the substrate binding site for YjhB is located at this central cavity, bound by transmembrane helices 1, 5, 7 and 8 (Figure 3.10). Importantly, this places the substrate binding site at the interface between the N- and C-terminal domains, consistent with the characteristic feature of the MFS (33).

In order to transport a sugar molecule against its electrochemical and/ or concentration gradient, symporters must couple this uphill movement with the translocation of a proton or ion down an electrochemical gradient (33). This process is typically driven by residues that can be protonated or deprotonated, such as Asp or Glu (64; 65). Among the homologs of YjhB, a conserved Asp on transmembrane helix 1 is one such residue, as mutagenesis of this residue completely abolishes transport (66). In YjhB, this key residue is identified as Asp27, while further structural analysis identified Asp30 and Glu120 as potential residues involved in proton/ion transfer. Bound by transmembrane helices 1, 3, 4 and 6, I propose that these residues are involved in the proton/ion translocation pore for YjhB (Figure 3.10).

Despite identifying a hypothetical location of the substrate binding site and proton/ion translocation pore, questions about the mechanism remain: 1) how do these residues create a proton/ion relay? 2) how is this relay coupled to the uphill movement and release of the substrate? and 3) could sialic acid be recognised by YjhB? It is important to note that further experimental and computational evidence, such as X-ray crystallography or molecular dynamic simulations are required here to quantitatively address these questions and define the ion dependency for YjhB. However, a hypothesis can still be made following further structural analysis using structural homologs and the *de novo* 3D model of YjhB.

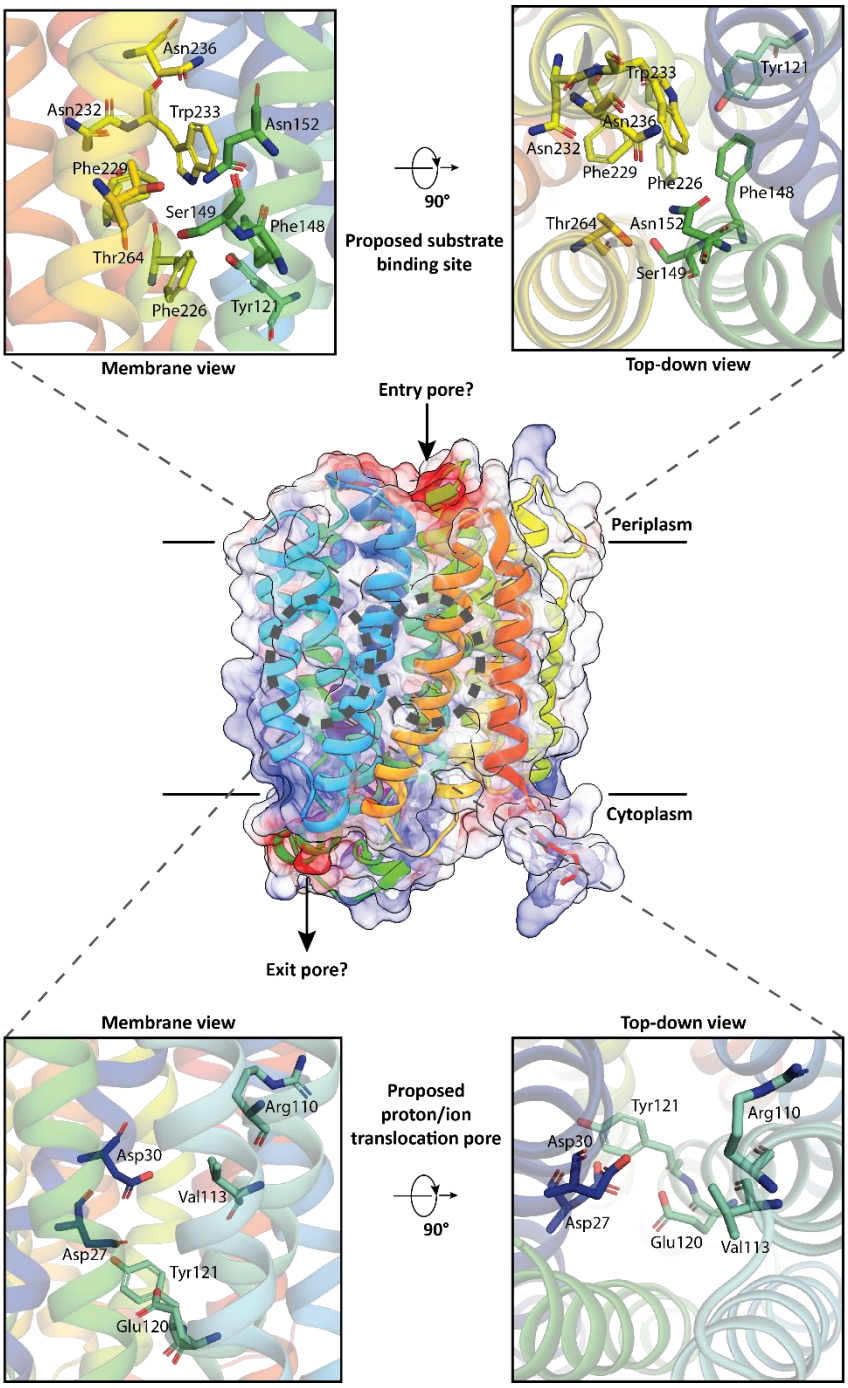


Figure 3.10 | Proposed substrate binding site and proton/ion translocation pore for *E. coli* YjhB. Based on structural analysis with closely related homologs, residues were identified that may be functionally important. The top panels highlight the aromatic and polar residues that may coordinate the preferred substrate at the proposed binding site. The bottom panels highlight the residues that may assist in a proton/ion relay through the translocation pore. The conserved Asp27 has been recognised to play this role among closely related structures (64; 66). The central diagram indicates the location of these proposed sites on the overall *de novo* model. Based on the electrostatic potential mapped to the surface of the model, entry and exit pores can be proposed on the basis of charge distribution. Figures were generated using PyMOL (67) and UCSF Chimera (68).

A basic residue located near the conserved Asp residue has been recognised to play an essential role in coupling the uphill movement of the substrate with proton translocation (64; 66). In YjhB, this is identified as Arg110, located on transmembrane helix 4 (Figure 3.10). In the absence of a proton, this basic residue is reported to form a salt bridge with the conserved Asp residue causing the transporter to switch to an outward-facing open conformation, where the substrate can now enter and bind in the central cavity (69). As an example, mutation of Asp128 in the lipophilic drug transporter LmrP significantly reduces its transport activity (70). It is hypothesised that protonation of the conserved Asp residue breaks this salt bridge causing a structural rearrangement within the N- and C-terminal domains and facilitating a switch to the inward-facing open conformation. As a result of this switch, translocation of the substrate and proton transfer between Asp and Glu residues is facilitated in what can be described as proton relay (69).

The conservation of the Asp residue in the translocation pore was originally identified as a characteristic feature of proton symporters, although in 2013, a study by Iancu and colleagues determined that the chemical environment within the translocation pore was also important (66). A sequence analysis revealed that while glucose transporter 2 contains this conserved Asp residue, it surprisingly functioned as a uniporter. Subsequently, modelling revealed that glucose transporter 2 had a Ser residue in the vicinity, which was able to form a hydrogen bond with the conserved Asp residue as opposed to a surrounding basic residue. As a result, the critical salt bridge was not formed. All other symporters in the study had a hydrophobic residue in this location (66). In YjhC, this hydrophobic residue is conserved (Val113), which together with the conserved Asp residue (Asp27) and the surrounding Arg110 suggests that YjhB could be a cation symporter.

Structural analysis also identified a key tyrosine residue, hypothesised to play a gating role in proton translocation in the phosphate transporter from *Piriformospora indica* (71)—this is conserved in YjhB, as Tyr121 (Figure 3.10). Specifically, when this tyrosine residue interacts with the substrate it leads to a small conformational change in transmembrane helix 4 of the phosphate transporter allowing protons along the proton relay to exit into the cytoplasm (71). To address if this tyrosine residue in YjhB is involved in the proton/ion relay, molecular dynamic simulations are required.

YjhB has been hypothesised to transport less common derivatives of sialic acid in bacteria (13). In consideration of its sequence identity to NanT, which is known to transport the common forms of sialic acid (72; 73), perhaps YjhB supports the role of NanT by expanding the substrate specificity for *E. coli* in the human host. To test this hypothesis, an *in vivo* assay using gene knockouts of the respective transporters, along with sialic acid derivatives should be performed in the future. Furthermore, while the residues identified in the substrate-binding site would support the binding of a sugar, no charged residue was observed. This contrasts with other sugar-proton symporters (10; 74), and *Proteus mirabilis*

SiaT; a sialic acid transporter within the sodium solute symporter family, which coordinates the carboxylate group of Neu5Ac through a salt bridge with Arg135 (61). The only reported sugar-proton symporter to not employ a charged residue in substrate binding is XylE (31). Therefore, perhaps YjhB functions to transport uncharged sialic acid derivatives, such as the lactone (Chapter One, Figure 1.1), which is the second most abundant sialic acid within the human gastrointestinal and respiratory tract (7; 13). Biologically, this would make sense as NanT can transport the acidic Neu5Ac and is likely to present specifically to other closely related acidic derivatives. If the lactone is transported, further processing was suggested through evidence of binding to YjhC (Manuscript, Chapter Two).

3.4 Chapter summary

This chapter investigated the previously uncharacterised YjhB membrane protein and addressed the hypothesis that the functional role of this transporter is to import the less common derivatives of sialic acid.

Using a protein-GFP fusion system, an overexpression protocol was developed for YjhB. This is a significant step towards ensuring a suitable amount of protein can be isolated for future structural and biophysical characterisation. Insight into the biological function of YjhB was explored through a *de novo* 3D model and *in silico* comparison between sequence homologs.

In brief, to summarise this *in silico* comparison the location of the substrate binding site and proton/ion translocation pore for YjhB have been proposed following interpretation among structural homologs. Bound by transmembrane helices 1, 5, 7 and 8, the proposed substrate binding site maintains a conservation of aromatic and polar residues to support the binding of a sugar-based substrate. Although, no charged residue was identified at this location, suggesting that YjhB does not transport sugar-based substrates with a net positive or negative charge. Considering the hypothesis towards the functional role of the *yjhBC* operon and the homology with NanR, this suggests that YjhB imports uncharged derivatives of sialic acid. An example could be the lactone derivative, identified as a promising lead for YjhC in Chapter Two. The conservation of acidic residues (Asp27, Asp30 and Glu120) and partners to form a key salt bridge (Asp27/Asp30 and Arg110) in the translocation pore suggest the role of YjhB as a cation symporter. Furthermore, a tyrosine residue (Tyr121) may play a gating role in the translocation mechanism.

Together, this characterisation completes the investigation into the uncharacterised *yjhBC* operon and has supplied data which enhances our understanding of sialic acid degradation in bacteria, providing a foundation to inform rational drug design to this viable target.

3.5 Chapter References

1. Pao, S. S., Paulsen, I. T., & Saier, M. H., Jr. (1998). Major facilitator superfamily. *Microbiology and Molecular Biology Reviews, 62*, 1-34.
2. Griffith, J. K., Baker, M. E., Rouch, D. A., Page, M. G., Skurray, R. A., Paulsen, I. T., Chater, K. F., Baldwin, S. A., & Henderson, P. J. (1992). Membrane transport proteins: implications of sequence comparisons. *Current Opinion in Cell Biology, 4*, 684-695.
3. Madej, M. G. (2014). Function, Structure, and Evolution of the Major Facilitator Superfamily: The LacY Manifesto. *Advances in Biology, 2014*, 20.
4. Shechter, E. (1986). Transports actifs secondaires. *Biochimie, 68*, 357-365.
5. Poolman, B., & Konings, W. N. (1993). Secondary solute transport in bacteria. *Biochimica et Biophysica Acta, 1183*, 5-39.
6. Vimr, E. R., Kalivoda, K. A., Deszo, E. L., & Steenbergen, S. M. (2004). Diversity of microbial sialic acid metabolism. *Microbiology and Molecular Biology Reviews, 68*, 132-153.
7. North, R. A., Horne, C. R., Davies, J. S., Remus, D. M., Muscroft-Taylor, A. C., Goyal, P., Wahlgren, W. Y., Ramaswamy, S., Friemann, R., & Dobson, R. C. J. (2018). "Just a spoonful of sugar...": import of sialic acid across bacterial cell membranes. *Biophysical Reviews, 10*, 219-227.
8. Saier, M. H., Jr. (2003). Tracing pathways of transport protein evolution. *Molecular Microbiology, 48*, 1145-1156.
9. Martinez, J., Steenbergen, S., & Vimr, E. (1995). Derived structure of the putative sialic acid transporter from *Escherichia coli* predicts a novel sugar permease domain. *Journal of Bacteriology, 177*, 6005-6010.
10. Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R., & Iwata, S. (2003). Structure and mechanism of the lactose permease of *Escherichia coli*. *Science, 301*, 610-615.
11. Huang, Y., Lemieux, M. J., Song, J., Auer, M., & Wang, D. N. (2003). Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science, 301*, 616-620.
12. Jiang, D., Zhao, Y., Wang, X., Fan, J., Heng, J., Liu, X., Feng, W., Kang, X., Huang, B., Liu, J., & Zhang, X. C. (2013). Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 14664.
13. Vimr, E. R. (2013). Unified theory of bacterial sialometabolism: how and why bacteria metabolise host sialic acids. *International Scholarly Research Notices Microbiology, 2013*, 816713.
14. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
15. Vimr, E. R., & Troy, F. A. (1985). Identification of an inducible catabolic system for sialic acids (nan) in *Escherichia coli*. *Journal of Bacteriology, 164*, 845-853.
16. Gordon, E., Horsefield, R., Swarts, H. G. P., De Pont, J. J. H. H. M., Neutze, R., & Snijder, A. (2008). Effective high-throughput overproduction of membrane proteins in *Escherichia coli*. *Protein expression and purification, 62*, 1-8.
17. Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli*. *Current Opinion in Biotechnology, 10*, 411-421.
18. Schlegel, S., Lofblom, J., Lee, C., Hjelm, A., Klepsch, M., Strous, M., Drew, D., Slotboom, D. J., & de Gier, J. W. (2012). Optimizing membrane protein overexpression in the *Escherichia coli* strain Lemo21(DE3). *Journal of Molecular Biology, 423*, 648-659.
19. Wagner, S., Klepsch, M. M., Schlegel, S., Appel, A., Draheim, R., Tarry, M., Hogbom, M., van Wijk, K. J., Slotboom, D. J., Persson, J. O., & de Gier, J. W. (2008). Tuning *Escherichia coli* for membrane protein overexpression. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 14371-14376.
20. Rosano, G. L., & Ceccarelli, E. A. (2014). Recombinant protein expression in *Escherichia coli*: advances and challenges. *Frontiers in Microbiology, 5*, 172.
21. Wagner, S., Bader, M. L., Drew, D., & de Gier, J. W. (2006). Rationalizing membrane protein overexpression. *Trends in Biotechnology, 24*, 364-371.
22. Giacalone, M. J., Gentile, A. M., Lovitt, B. T., Berkley, N. L., Gunderson, C. W., & Surber, M. W. (2006). Toxic protein expression in *Escherichia coli* using a rhamnose-based tightly regulated and tunable promoter system. *Biotechniques, 40*, 355-364.
23. Hsieh, J. M., Besserer, G. M., Madej, M. G., Bui, H. Q., Kwon, S., & Abramson, J. (2010). Bridging the gap: a GFP-based strategy for overexpression and purification of membrane proteins with intra and extracellular C-termini. *Protein Science, 19*, 868-880.

24. Drew, D., Lerch, M., Kunji, E., Slotboom, D. J., & de Gier, J. W. (2006). Optimisation of membrane protein overexpression and purification using GFP fusions. *Nature Methods, 3*, 303-313.
25. Rath, A., Glibowicka, M., Nadeau, V. G., Chen, G., & Deber, C. M. (2009). Detergent binding explains anomalous SDS-PAGE migration of membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 1760-1765.
26. Kawate, T., & Gouaux, E. (2006). Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure, 14*, 673-681.
27. Whipp, M. J., Camakaris, H., & Pittard, A. J. (1998). Cloning and analysis of the *shiA g*ene, which encodes the shikimate transport system of *Escherichia coli* K-12. *Gene, 209*, 185-192.
28. Xu, Y., Gao, X., Wang, S. H., Liu, H., Williams, P. A., & Zhou, N. Y. (2012). MhbT is a specific transporter for 3-hydroxybenzoate uptake by Gram-negative bacteria. *Applied Environmental Microbiology, 78*, 6113-6120.
29. Seol, W., & Shatkin, A. J. (1993). Membrane topology model of *Escherichia coli* alpha-ketoglutarate permease by phoA fusion analysis. *Journal of Bacteriology, 175*, 565-567.
30. Rodionov, D. A., Li, X., Rodionova, I. A., Yang, C., Sorci, L., Dervyn, E., Martynowski, D., Zhang, H., Gelfand, M. S., & Osterman, A. L. (2008). Transcriptional regulation of NAD metabolism in bacteria: genomic reconstruction of NiaR (YrxA) regulon. *Nucleic Acids Research, 36*, 2032-2046.
31. Sun, L., Zeng, X., Yan, C., Sun, X., Gong, X., Rao, Y., & Yan, N. (2012). Crystal structure of a bacterial homologue of glucose transporters GLUT1-4. *Nature, 490*, 361-366.
32. Tsirigos, K. D., Peters, C., Shu, N., Käll, L., & Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Research, 43*, W401-W407.
33. Quistgaard, E. M., Low, C., Guettou, F., & Nordlund, P. (2016). Understanding transport by the major facilitator superfamily (MFS): structures pave the way. *Nature Reviews Molecular Cell Biology, 17*, 123-132.
34. Tan, S., Tan, H. T., & Chung, M. C. (2008). Membrane proteins and membrane proteomics. *Proteomics, 8*, 3924-3932.
35. Rawlings, A. E. (2016). Membrane proteins: always an insoluble problem? *Biochemical Society Transactions, 44*, 790-795.
36. Cheng, Y. (2018). Membrane protein structural biology in the era of single particle cryo-EM. *Current Opinion in Structural Biology, 52*, 58-63.
37. Kuhlbrandt, W. (2014). Biochemistry. The resolution revolution. *Science, 343*, 1443-1444.
38. Thonghin, N., Kargas, V., Clews, J., & Ford, R. C. (2018). Cryo-electron microscopy of membrane proteins. *Methods, 147*, 176-186.
39. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Jr., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation, 13*, 3031-3048.
40. Wang, S., Li, W., Zhang, R., Liu, S., & Xu, J. (2016). CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Research, 44*, W361-366.
41. Kopp, J., & Schwede, T. (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics, 5*, 405-416.
42. Bonneau, R., & Baker, D. (2001). *Ab initio* protein structure prediction: progress and prospects. *Annual Review of Biophysics and Biomolecular Structure, 30*, 173-189.
43. Di Lena, P., Nagata, K., & Baldi, P. (2012). Deep architectures for protein contact map prediction. *Bioinformatics, 28*, 2449-2457.
44. de Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics, 14*, 249-261.
45. Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate *De Novo* Prediction of Protein Contact Map by Ultra-Deep Learning Model. *Plos Computational Biology, 13*, e1005324.
46. Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology, 268*, 209-225.
47. Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research, 32*, W526-531.
48. Alford, R. F., Koehler Leman, J., Weitzner, B. D., Duran, A. M., Tilley, D. C., Elazar, A., & Gray, J. J. (2015). An Integrated Framework Advancing Membrane Protein Modeling and Design. *Plos Computational Biology, 11*, e1004398.

49. Park, H., Bradley, P., Greisen, P., Jr., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D., & DiMaio, F. (2016). Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation, 12*, 6201-6212.
50. Barth, P., Schonbrun, J., & Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 15682-15687.
51. Lam, S. D., Das, S., Sillitoe, I., & Orengo, C. (2017). An overview of comparative modelling and resources dedicated to large-scale modelling of genome sequences. *Acta Crystallographica Section D Structural Biology, 73*, 628-640.
52. Wang, S., Fei, S., Wang, Z., Li, Y., Xu, J., Zhao, F., & Gao, X. (2019). PredMP: a web server for *de novo* prediction and visualization of membrane proteins. *Bioinformatics, 35*, 691-693.
53. Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research, 44*, W430-435.
54. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., & Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D, 54*, 905-921.
55. Lomize, A. L., Pogozheva, I. D., Lomize, M. A., & Mosberg, H. I. (2006). Positioning of proteins in membranes: a computational approach. *Protein science : a publication of the Protein Society, 15*, 1318-1333.
56. Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., Krasny, R., Wei, G. W., Holst, M. J., McCammon, J. A., & Baker, N. A. (2018). Improvements to the APBS biomolecular solvation software suite. *Protein Science, 27*, 112-128.
57. Henderson, P. J., & Maiden, M. C. (1990). Homologous sugar transport proteins in *Escherichia coli* and their relatives in both prokaryotes and eukaryotes. *Philosophical Transactions of the Royal Society B-Biological Sciences, 326*, 391-410.
58. Maiden, M. C., Davis, E. O., Baldwin, S. A., Moore, D. C., & Henderson, P. J. (1987). Mammalian and bacterial sugar transport proteins are homologous. *Nature, 325*, 641-643.
59. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soeding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology, 7*.
60. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research, 44*, W344-W350.
61. Wahlgren, W. Y., Dunevall, E., North, R. A., Paz, A., Scalise, M., Bisignano, P., Bengtsson-Palme, J., Goyal, P., Claesson, E., Caing-Carlsson, R., Andersson, R., Beis, K., Nilsson, U. J., Farewell, A., Pochini, L., Indiveri, C., Grabe, M., Dobson, R. C. J., Abramson, J., Ramaswamy, S., & Friemann, R. (2018). Substrate-bound outward-open structure of a Na(+)-coupled sialic acid symporter reveals a new Na(+) site. *Nature Communications, 9*, 1753.
62. Faham, S., Watanabe, A., Besserer, G. M., Cascio, D., Specht, A., Hirayama, B. A., Wright, E. M., & Abramson, J. (2008). The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na+/sugar symport. *Science, 321*, 810-814.
63. Perez, C., Faust, B., Mehdipour, A. R., Francesconi, K. A., Forrest, L. R., & Ziegler, C. (2014). Substrate-bound outward-open state of the betaine transporter BetP provides insights into Na+ coupling. *Nature Communications, 5*, 4231.
64. Wisedchaisri, G., Park, M. S., Iadanza, M. G., Zheng, H., & Gonen, T. (2014). Proton-coupled sugar transport in the prototypical major facilitator superfamily protein XylE. *Nature Communications, 5*, 4521.
65. Zheng, H., Wisedchaisri, G., & Gonen, T. (2013). Crystal structure of a nitrate/nitrite exchanger. *Nature, 497*, 647-651.
66. Iancu, C. V., Zamoon, J., Woo, S. B., Aleshin, A., & Choe, J. Y. (2013). Crystal structure of a glucose/H+ symporter and its mechanism of action. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 17862-17867.
67. DeLano, W. L. (2002). The PyMOL Molecular Graphics Program. San Carlos: Delano Scientific.
68. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry, 25*, 1605-1612.

69. Zhang, X. C., Zhao, Y., Heng, J., & Jiang, D. (2015). Energy coupling mechanisms of MFS transporters. *Protein science : a publication of the Protein Society, 24*, 1560-1579.
70. Gbaguidi, B., Mazurkiewicz, P., Konings, W. N., Driessen, A. J., Ruysschaert, J. M., & Vigano, C. (2004). Proton motive force mediates a reorientation of the cytosolic domains of the multidrug transporter LmrP. *Cellular and Molecualr Life Science, 61*, 2646-2657.
71. Pedersen, B. P., Kumar, H., Waight, A. B., Risenmay, A. J., Roe-Zurz, Z., Chau, B. H., Schlessinger, A., Bonomi, M., Harries, W., Sali, A., Johri, A. K., & Stroud, R. M. (2013). Crystal structure of a eukaryotic phosphate transporter. *Nature, 496*, 533-536.
72. Nilsson, I., Prathapam, R., Grove, K., Lapointe, G., & Six, D. A. (2018). The sialic acid transporter NanT is necessary and sufficient for uptake of 3-deoxy-d-manno-oct-2-ulosonic acid (Kdo) and its azido analog in *Escherichia coli*. *Molecular Microbiology, 110*, 204-218.
73. Mulligan, C., Geertsma, E. R., Severi, E., Kelly, D. J., Poolman, B., & Thomas, G. H. (2009). The substrate-binding protein imposes directionality on an electrochemical sodium gradient-driven TRAP transporter. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 1778-1783.
74. Dang, S., Sun, L., Huang, Y., Lu, F., Liu, Y., Gong, H., Wang, J., & Yan, N. (2010). Structure of a fucose transporter in an outward-open conformation. *Nature, 467*, 734-738.

Chapter Four

Towards a structural and functional understanding of the RpiR-type sialoregulator from *Streptococcus pneumoniae*

4.1 Introduction

4.1.1 RpiR family of transcriptional regulators

The RpiR family of transcriptional regulators regulate the genes involved in sugar catabolic pathways (1) and is found in both Gram-positive bacteria, such as *Bacillus subtilis*, *Staphylococcus aureus* and *Streptococcus pneumoniae* and Gram-negative bacteria, such as *Escherichia coli*, *Haemophilus influenzae* and *Pseudomonas spp.* (2-6). These catabolic pathways include the metabolism of glucose, maltose and ribose sugars, the pentose phosphate pathway, inositol catabolism, *N*-acetylmuramic acid catabolism and *N*-acetylneuraminic acid catabolism (1; 2; 7; 8).

Within the cell, RpiR transcriptional regulators operate as both gene activators and gene repressors for these catabolic enzymes (7). The first member to be identified for this family was the transcriptional regulator RpiR, which acts as a transcriptional repressor to negatively control the gene expression of *rpiB* in *E. coli* (3). This gene encodes an isomerase that catalyses the isomerisation of ribose-5-phosphate and ribulose-5-phosphate as part of the pentose phosphate pathway (3; 7).

Members of the RpiR family of transcriptional regulators are structurally comprised of an N-terminal DNA-binding domain, which contains a helix-turn-helix motif, and a C-terminal sugar isomerase domain (7; 9). The helix-turn-helix motif is one of the best characterised DNA-binding motifs amongst bacterial transcriptional regulators and is defined by a tight tri-helical structure, where the third α-helix is often referred to as the "recognition helix", as it interacts directly with the major groove of the DNA (10) (Figure 4.1A). This is achieved primarily through electrostatic interactions, *via* positively-charged amino acid residues that interact with the DNA bases that form the recognition sequence for the regulator (11). In addition, polar interactions primarily with the DNA backbone provide further stability for the overall structure.

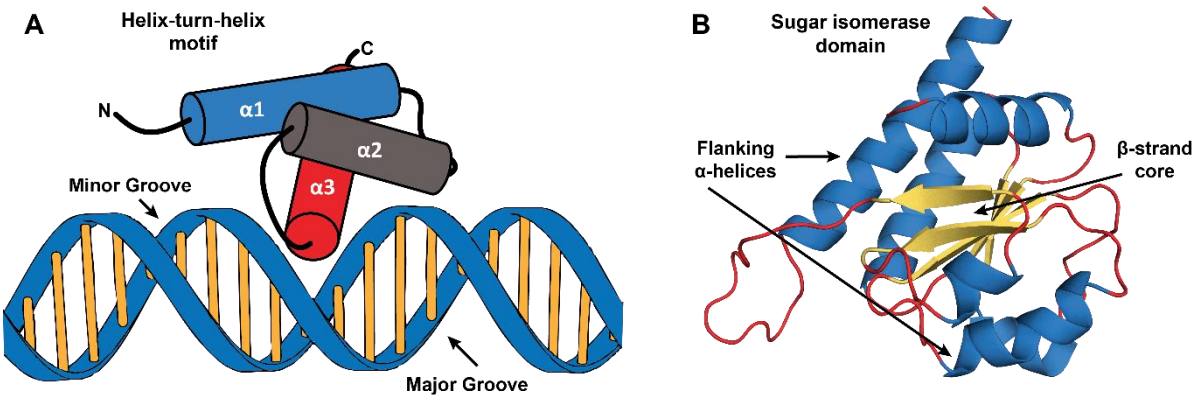


Figure 4.1 | Cartoon overview of the helix-turn-helix motif and the sugar isomerase domain. **A)** The helix-turn-helix motif is one of the best characterised DNA binding motifs amongst bacterial transcriptional regulators. Structurally, this motif is comprised of a tight tri-helical architecture, where the third α-helix (α3 in red) is often referred to as the 'recognition helix', as it interacts directly with the major groove of the DNA primarily via electrostatic interaction. **B)** The sugar isomerase domain is characterised by an alpha-beta architecture comprising multiple α-helices (blue) flanking a β-strand (orange) core. This cartoon representation is taken from the *E. coli* arabinose-5-phosphate isomerase (PDB ID 2XHZ) as a representative structure.

The sugar isomerase domain is represented across eukaryotic, archaea and prokaryotic proteins and usually binds a phosphorylated sugar (9). Despite sharing a low sequence identity, structurally these domains share an alpha-beta architecture that is characterised by multiple α-helices surrounding a β-strand core (Figure 4.1B) (12). Commonly, these proteins serve a catalytic role and function as an isomerase (9), however, in the RpiR family the sugar isomerase domain has been co-opted into an allosteric role in the regulatory mechanism (1; 13). Here, RpiR-mediated regulation can orchestrate the metabolism of carbohydrate nutrients by sensing the concentration of phosphosugars within the cellular environment (14). Upon binding, a conformational change is propagated through the regulator, attenuating the specific DNA interaction or binding affinity, which causes the regulator to dissociate (2; 15). As a result, gene expression is induced for the metabolism of the respective carbohydrate (2). The phosphosugar that induces this allosteric response is referred to as the effector molecule.

There are only three crystal structures of RpiR-type transcriptional regulators. One of these is a regulator of sialic acid catabolism in *Vibro vulnificus* (PDB entry 4IVN), which is discussed below, while the other two are solved as part of structural genomics project and are therefore not published, where they contain either the N-terminal or C-terminal domains only (PDB ID 31WF and 3SHO, respectively). Consequently, at a structural level the RpiR family of transcriptional regulators is poorly understood.

4.1.2 The RpiR-type sialoregulators

Given the importance of sialic acid to the survival and persistence of bacterial pathogens (16; 17), it is unsurprising that its catabolic pathway is tightly regulated to ensure this alternative nutrient is utilised to its full potential when present in the surrounding environment. In response to this availability, the transcriptional regulator NanR orchestrates the gene expression of the sialic acid catabolic machinery in order to utilise the nutrient source (13; 14). Largely placed in the RpiR-family of transcriptional regulators, NanR has been identified in both Gram-positive and Gram-negative bacterial pathogens including *Clostridium perfringens* (8), *H. influenzae* (5), *S. aureus* (6), *S. pneumoniae* (4; 18) and *V. vulnificu*s (13; 19). Given these transcriptional regulators are involved in the gene regulation of sialic acid catabolism, they are referred to as 'sialoregulators' (20).

Work to date has focused on distinguishing the genes that are controlled by these sialoregulators and identifying the consensus nucleotide sequence that is recognised to facilitate regulation (4-6; 8; 13; 19). This research demonstrates that NanR collectively regulates the gene expression of the *nanA*, *nanK* and *nanE* genes, which encode the core catabolic enzymes, *N*-acetylneuraminate lyase, *N*-acetylmannosamine kinase and *N*-acetylmannosamine 6-phosphate 2-epimerase (4; 8; 13; 18). In addition, NanR also regulates gene expression for various other proteins involved in the import or processing of sialic acid, including: 1) the specific transporters of sialic acid such as the secondary transporter NanT or the TRAP transporters (8); 2) several uncharacterised proteins, such as the putative oxidoreductase YjhC, which is investigated in Chapter Two; and 3) a sialidase (NeuA) used for scavenging sialic acid from glycoconjugates (4; 21). Collectively, the occurrence of the *nanA*, *nanK* and *nanE* genes are highly conserved among bacteria that can catabolise sialic acid, while the genes encoding the transport machinery or enzymes required to further process sialic acid can change depending on the species.

Secondly, various degradation products of the sialic acid catabolic pathway have been screened in order to identify potential effector molecules that induce the *in vivo* expression of the proteins required to degrade sialic acid. Gualdi *et al.* (2012) demonstrate *in vivo* that the *S. pneumoniae* regulator is responsive to the pathway intermediates *N*-acetylmannosamine, or *N*-acetylmannosamine-6-phosphate (18). In addition, Olson et al. (2013) demonstrate that in the presence of the phosphorylated pathway intermediate *N*-acetylmannosamine-6-phosphate, the *S. aureus* regulator exhibits a loss of DNA-binding affinity (6). This attenuation of DNA-binding activity in the presence of *N*-acetylmannosamine-6-phosphate is also observed in NanR from *V. vulnificus* (13), while glucosamine-6-phosphate is implicated as the effector molecule in *H. influenzae* NanR (5). Thus, the RpiR-type

sialoregulators are largely responsive to phosphorylated pathway intermediates, a feature that is facilitated by the sugar isomerase domain.

The structural characterisation of the RpiR-type sialoregulators is limited to NanR from *V. vulnificus*, where the crystal structure has been solved to 1.9 Å in complex with its effector, *N*-acetylmannosamine-6-phosphate (PDB ID 4IVN). The authors demonstrate that this NanR forms a dimeric assembly, where DNA is hypothesised to bind between the monomers in an arched tunnel-like cavity, mediated by positively charged residues (13). In addition, this structure provides the first insight into the coordination of the effector molecule within the sugar isomerase domain, which is primarily mediated *via* electrostatic interactions and the amino acid residue H163. However, as discussed in Chapter One, the dimer interface of their proposed 'active form' is unconvincing, as it is formed by only four hydrogen bonds and buries less than 3% of the accessible surface area (Figure 4.2). Thus, the true oligomeric assembly of the RpiR-type sialoregulators, as well as the molecular details that define the protein-DNA interaction are open questions. This is investigated in this Chapter.

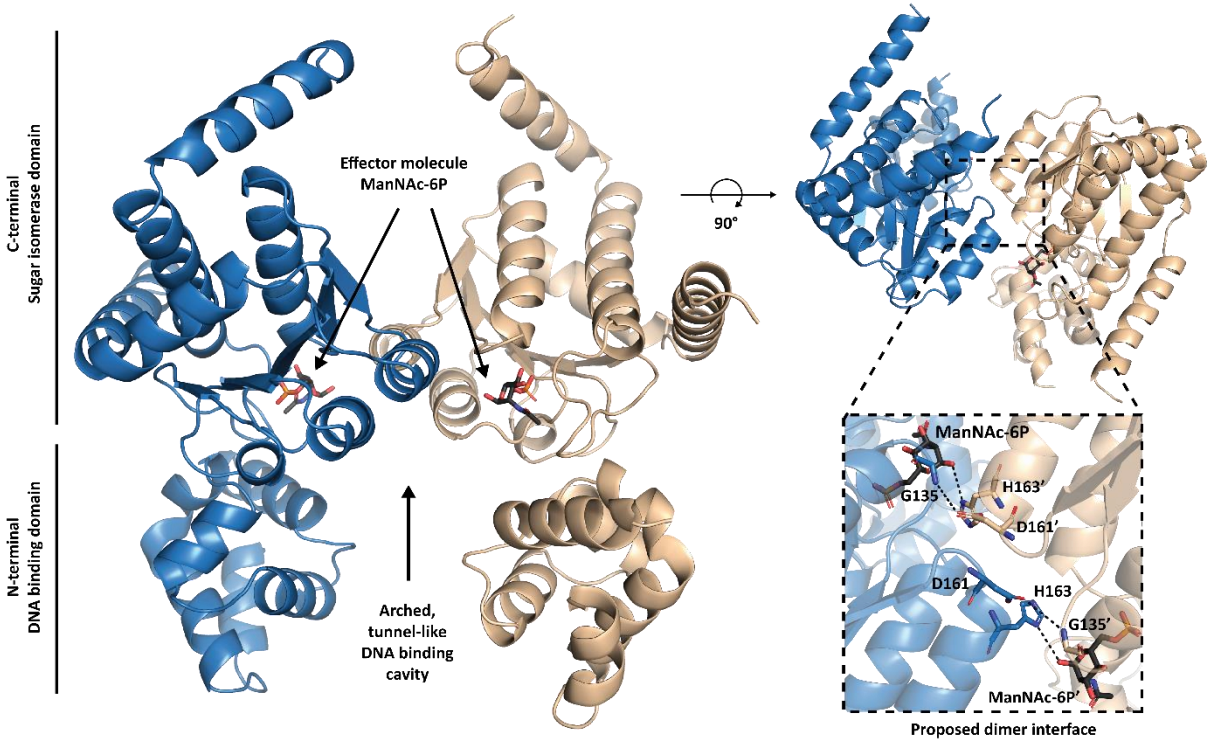


Figure 4.2 | Crystal structure of the *V. vulnificus* NanR. The left panel presents the proposed 'active' dimeric assembly for *V. vulnificus* RpiR-type NanR in its 'active state' in complex with the *N*-acetylmannosamine-6-phopshate (ManNAc-6P) effector molecule (sticks). Highlighted is the arched, tunnel-like DNA-binding cavity. The right panel presents a top view of the transcriptional regulator, where the insert highlights the four hydrogen bonds that form the dimeric interface—these are formed between residues, D161-G135 and H163-ManNAc-6P. This figure was generated using PyMOL and the PDB ID 4IVN (13).

4.1.3 *S. pneumoniae* NanR

S. pneumoniae is a low-GC Gram-positive bacterial pathogen of the human respiratory system and is a major cause of pneumonia, meningitis, bacteraemia and ear infections (otitis media) (22). To establish infection within the human host, *S. pneumoniae* have evolved the ability to utilise sialic acid as an alternative source of nutrition (4; 18; 23). In addition, *S. pneumoniae* is known to express surface-associated sialidases that function to cleave sialic acid from the terminal of glycoconjugates. Being able to catabolise and recycle sialic acids assists colonisation within the host, increases pathogenicity and provides a fitness benefit over pathogens who rely on endogenously expressed sialidases of the host to scavenge sialic acid (18; 21; 24; 25).

The gene expression of the proteins required to utilise sialic acid in *S. pneumoniae* is under control of the RpiR-type transcriptional regulator NanR. Here, NanR binds to two different operons, *nan* operon I, and *nan* operon II, while additionally regulating the expression of the sialidase gene, *nanA*—these genes are clustered within the genome (4). *nan* operon I is comprised of 10 genes that encode the essential catabolic enzymes NanA, NanK and NanE, as well as various hypothetical proteins and a phosphotransferase system. Conversely, *nan* operon II is comprised of six primary hypothetical proteins and a putative oxidoreductase. This oxidoreductase is placed in the Gfo/Idh/MocA family, as is YjhC, which is investigated in Chapter Two. The recognition sequence for these operons has been identified (4; 18) and is presented in Figure 4.3, where it is compared with other Streptococci strains from the RegPrecise database (26). Overall, this DNA recognition sequence for NanR is highly conserved among Streptococci. It is unknown if these sequences will form hairpin structures *in vivo*, on the basis of the base complementarity between positions 3 to 7 and 12 to 16 (Figure 4.3).

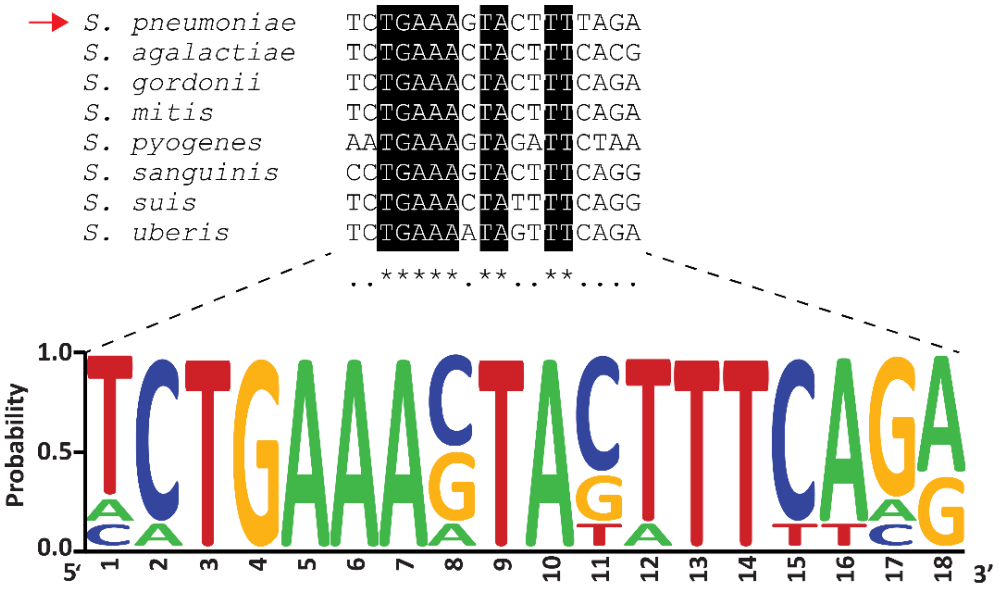


Figure 4.3 | The DNA recognition sequence for NanR within Streptococci. Eight different DNA recognition sequences are presented, including *S. pneumoniae* (labelled with red arrow). The regions of conservation are highlighted in white text with a black background. Combined, these sequences were used to generate a sequence logo using WebLogo (27). The height of each base pair is representative of its probability or conservation among Streptococci strains.

4.2 Chapter overview

This chapter presents structural and functional studies of the RpiR-type NanR from the Gram-positive human pathogen, *S. pneumoniae*, a common inhabitant of the human respiratory tract and a significant cause of infection and disease (22; 28). The overall aim of this chapter is to define the biologically relevant quaternary structure and characterise the molecular details that govern the protein-DNA interaction.

The results demonstrate that *S. pneumoniae* NanR binds the phosphorylated sugar *N*-acetylmannosamine-6-phosphate, which is consistent with its implication as an effector molecule *in vivo* (18). A thorough investigation into the quaternary structure, stoichiometry of the protein-DNA hetero-complex and binding kinetics is presented using analytical ultracentrifugation. This is supported by solution scattering data, where additionally a multiphase *ab initio* reconstruction provides insight into the orientation of the protein and DNA within the hetero-complex. Although X-ray diffraction data was not suitable for structure determination, the crystals obtained are appropriate for future optimisation. Combined, these experiments enhance our understanding of how RpiR-type sialoregulators control gene expression.

4.3 Results and Discussion

4.3.1 Cloning, expression and purification of *S. pneumoniae* NanR

The *nanR* gene encoding the sialoregulator from *S. pneumoniae* was synthesised commercially in the cloning vector designated pUC57 with the restriction sites *Bam*HI and *Hind*III. To assist overexpression of the construct, the *nanR* gene was codon optimised for *E. coli*—the recombinant expression host. Initially, in order to use an N-terminal poly-Histidine tag (His-tag), the *nanR* gene was sub-cloned into the pET30ΔSE expression vector. However, the His-tag exhibited poor binding affinity during immobilised-metal affinity chromatography, with a significant portion eluting in the flow-through or wash. Moreover, no protein-DNA interaction was observed using this construct in preliminary DNA binding experiments and therefore the use of a His-Tag was not suitable for this system. To rectify this, PCR and In-Fusion cloning was employed to utilise the restriction sites *Nde*I and *Hind*III and sub-clone the *nanR* gene back into the pET30ΔSE expression vector, without the incorporation of the N-terminal His-tag.

S. pneumoniae NanR was successfully purified using a four-step procedure, employed over two days: ammonium sulphate precipitation, anion exchange chromatography, heparin chromatography and size-exclusion chromatography (Chapter Eight, Section 8.6.2.2). Following protein preparation, an initial ammonium sulphate precipitation step was utilised to remove nucleic acid contamination without the use of DNase, which may have affected DNA-binding studies following purification (29). This was verified by assessing the ratio of absorbance at 260/280 nm on a NanoDrop Spectrophotometer, where a ratio of 0.61 was obtained following this initial purification step. This ratio is consistent with a protein spectrum that is free of nucleic acid contamination. Without this step a 260/280 nm ratio at ~1.0 towards a value of 2.0 was obtained, indicative of increasing nucleic acid contamination (29). SDS-PAGE analysis following this procedure indicated that most of the protein was retained, thus while this is effective at removing nucleic acid it offers little overall purification. However, as characterising the protein-DNA interaction is a major aim of this study, this preliminary step was necessary. Prior to subjecting the protein to anion exchange chromatography, the sample was buffer exchanged by dialysis to ensure efficient binding to the column.

Following elution from the anion exchange column (Figure 4.4A), the protein sample was subjected to heparin chromatography. Here, the structure and negative charge of heparin allows it to mimic the polyanionic structure of DNA and serve as an affinity ligand (30). Following elution, NanR was primarily homogenous, where the flow-through contained the remaining population of protein (Figure 4.4B).

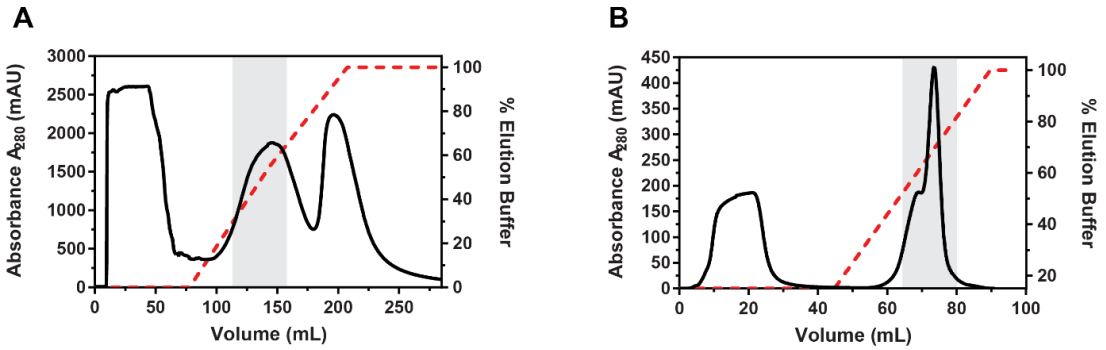


Figure 4.4 | A typical purification of *S. pneumoniae* NanR. A) A typical chromatogram from anion exchange chromatography. **B)** A typical chromatogram from heparin chromatography. The column was equilibrated with 150 mM NaCl as the recombinant protein was deemed unstable at lower salt concentrations. The red line highlights the concentration of elution buffer used to elute the recombinant protein, while the grey box highlights the fractionation range that was pooled.

As a final purification step, the fractions pooled from heparin chromatography were subjected to size-exclusion chromatography, where initially, the column was equilibrated with a buffer containing 150 mM NaCl. However, at this salt concentration the protein eluted as a very broad peak across a retention volume of 80 mL. This observation is indicative of non-specific ionic interactions with the Superdex matrix. To fix this, I increased the ionic strength of the buffer by using a salt concentration of 300 mM, where following equilibration with this new buffer, NanR eluted as a much tighter, single elution peak across a retention volume of approximately 10 mL (Figure 4.5A). The increasing purity of the recombinant protein from each chromatography technique is shown in Figure 4.5B. The final purity following size-exclusion chromatography was estimated to be approximately 95%, as highlighted by the single band (Figure 4.5B). Moreover, the theoretical monomeric molar mass of 32, 671 Da was verified by electrospray ionisation mass spectrophotometry, providing a molar mass of 32, 529 Da. This small discrepancy is consistent with the loss of the terminal methionine (149 Da). The final yield of *S. pneumoniae* NanR was approximately 20 to 30 mg per litre of culture.

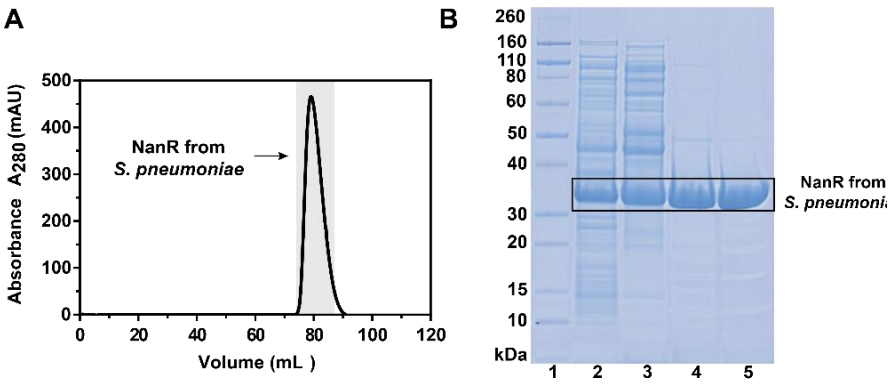


Figure 4.5 | Purification of *S. pneumoniae* NanR. A) A typical chromatogram from size-exclusion chromatography. The grey box highlights the fractionation range that was pooled and flash-frozen for future structural and functional characterisation. **B)** SDS-PAGE analysis showing the increasing purity following each purification step. Lane 1, Protein ladder (kDa); lane 2, crude lysate; lane 3, pooled fractions from anion-exchange chromatography; lane 4, pooled fractions from heparin chromatography; and lane 5, pooled fractions following size-exclusion chromatography.

4.3.2 The thermal stability of *S. pneumoniae* NanR

Three effector molecules for *S. pneumoniae* NanR have been identified, following *in vivo* growth studies reported in the literature (18; 31). These include Neu5Ac, and two metabolites of the sialic acid degradation pathway: *N*-acetylmannosamine; and *N*-acetylmannosamine-6-phosphate. To verify these effectors bind NanR *in vitro*, differential scanning fluorimetry experiments were used to detect changes in the melting temperature of the protein, in the presence and absence of the molecules (Chapter Eight, Section 8.6.2.3). Effector binding can be evaluated based upon the changes in melting temperature, as such a positive or negative thermal shift can be used to infer whether the effector molecule is binding the sialoregulator.

Initially, the thermal stability was measured for NanR in the absence of any effector molecule, demonstrating a melting temperature of 48.1 ± 0.1 °C (Figure 4.6). To assess the thermal stability of NanR in the presence of potential effector molecules, a total of six metabolites of the sialic acid degradation pathway were used, including: Neu5Ac; *N*-acetylmannosamine (ManNAc); *N*-acetylmannosamine-6-phosphate (ManNAc-6P); *N*-acetylglucosamine (GlcNAc); and *N*-acetylglucosamine-6-phosphate (GlcNAc-6P). A significant increase in thermal stability of 9.4 °C was only observed in the presence of ManNAc-6P (Figure 4.6). This provides strong evidence of a binding event for ManNAc-6P, supporting its implication as an effector molecule (18) and is consistent with the observation of RpiR family members generally binding phosphorylated end-products of their respective catabolic pathways (9). However, no significant increase in thermal stability was observed for Neu5Ac

or *N*-acetylmannosamine, contradicting the *in vivo* growth experiments in the literature (4; 18). Perhaps *in vivo* this interaction is either transient, or the result of degradation to ManNAc-6P when grown on Neu5Ac and ManNAc, both upstream metabolites in the sialic acid catabolic pathway (Chapter One, Figure 1.7). All other pathway metabolites presented a thermal shift of less than 1.0 °C, which was deemed insignificant.

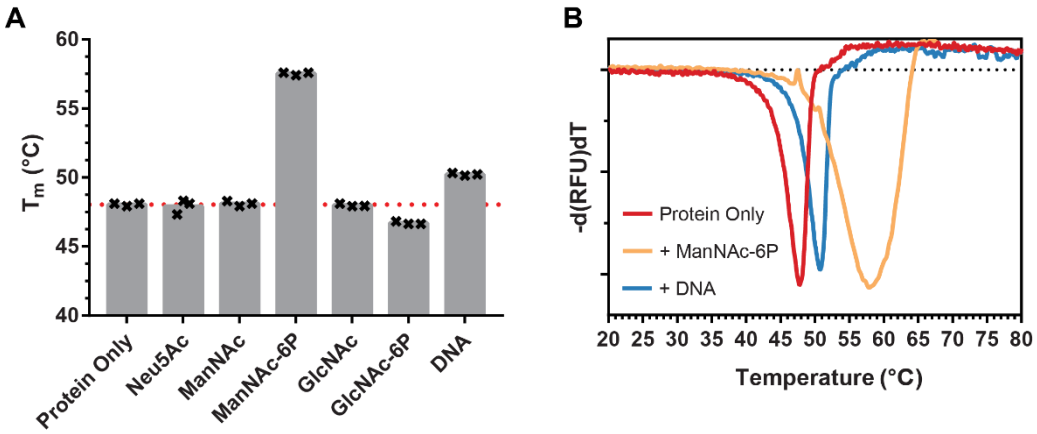


Figure 4.6 | Thermal stability of *S. pneumoniae* NanR in the presence and absence of potential effectors molecules. **A)** The mean melting temperatures determined for *S. pneumoniae* NanR is presented (bar). The melting temperature (T_m) from each triplicate measurement is shown (black cross). The red line highlights the T_m for the protein only. **B)** The thermal stability shift presented as a derivative plot. Here, the T_m is denoted as the lowest part of the curve. The conditions that demonstrated a significant increase in thermal stability relative to the protein by itself are shown.

Lastly, to assess if DNA provided a change in thermal stability, the protein was also incubated with the respective DNA recognition sequence (Chapter Eight, Table 8.4). Here, an increase of 2.1 °C was observed, providing evidence of DNA binding. DNA-binding activity is investigated in further detail later in the chapter. Together, the results from these thermal shift experiments are summarised in Table 4.1.

Table 4.1 | Summary of all T_m, where significant thermal shifts are highlighted in bold.

Molecule	T_m (°C)
	S. pneumoniae NanR
Protein only	48.1 ± 0.1
+ Neu5Ac	47.9 ± 0.5
+ ManNAc	48.1 ± 0.2
+ ManNAc-6P	**57.5 ± 0.1**
+ GlcNAc	48.0 ± 0.1
+ GlcNAc-6P	46.71 ± 0.1
+ DNA Recognition Sequence	**50.2 ± 0.1**

4.3.3 Sequence comparison with other RpiR-type sialoregulators

To elucidate potential amino acid residues that may be functionally important for the binding of the effector and interaction with DNA, a multiple sequence alignment was generated between *S. pneumoniae* NanR and the RpiR-type sialoregulators from *C. perfringens*, *H. influenzae*, *S. aureus*, and *V. vulnificus* using Clustal Omega (32). By comparing the *V. vulnificus* crystal structure (PDB ID 4IVN), potential amino acid residues that are functionally important were assigned based on conservation between sequences. These proteins share a sequence identity of 20-30%.

Within the N-terminal DNA-binding domain, the majority of amino acid residues that show high sequence conservation have positively-charged side chains or polar uncharged side chains (Figure 4.7, highlighted red). Within the *V. vulnificus* structure, these residues are primarily clustered together on the surface of the protein where they are poised for electrostatic or polar interaction with the DNA base pairs and backbone. A combination of arginine, lysine and threonine side chains is a common feature in the helix-turn-helix motif (10; 33). In addition, other conserved amino acid residues or chemically similar amino acid residues likely play a structural role, maintaining the correct fold of the helix-turn-helix motif and overall N-terminal domain.

Conversely, within the C-terminal sugar isomerase domain, the amino acid residues that coordinate the phosphate oxygens atoms of the effector molecule *N*-acetylmannosamine-6-phosphate in *V. vulnificus* are highly conserved (Figure 4.7, highlighted in blue; residues S182, S184 T187 in *V. vulnificus* sequence) (13). This amino acid conservation is consistent with the evidence from differential scanning fluorimetry experiments where only a phosphorylated sugar provided a significant increase in thermal stability (Section 3.3.3). In addition, other amino acids highlighted in Figure 4.7 provide additional coordination to the effector molecule, primarily through a water-mediated hydrogen-bonding network. In particular, proline at position 231 in *V. vulnificus* is observed to form a hydrogen bond with the carbonyl oxygen of the *N*-acetyl moiety in *N*-acetylmannosamine-6-phosphate (13). This proline is conserved in *S. pneumoniae* NanR at position 227. Conversely, *H. influenzae* NanR, which is predicted to bind glucosamine-6-phosphate (5) has a lysine at this position. As glucosamine-6-phosphate lacks the *N*-acetyl moiety, this feature may explain the differences in effector-binding specificity.

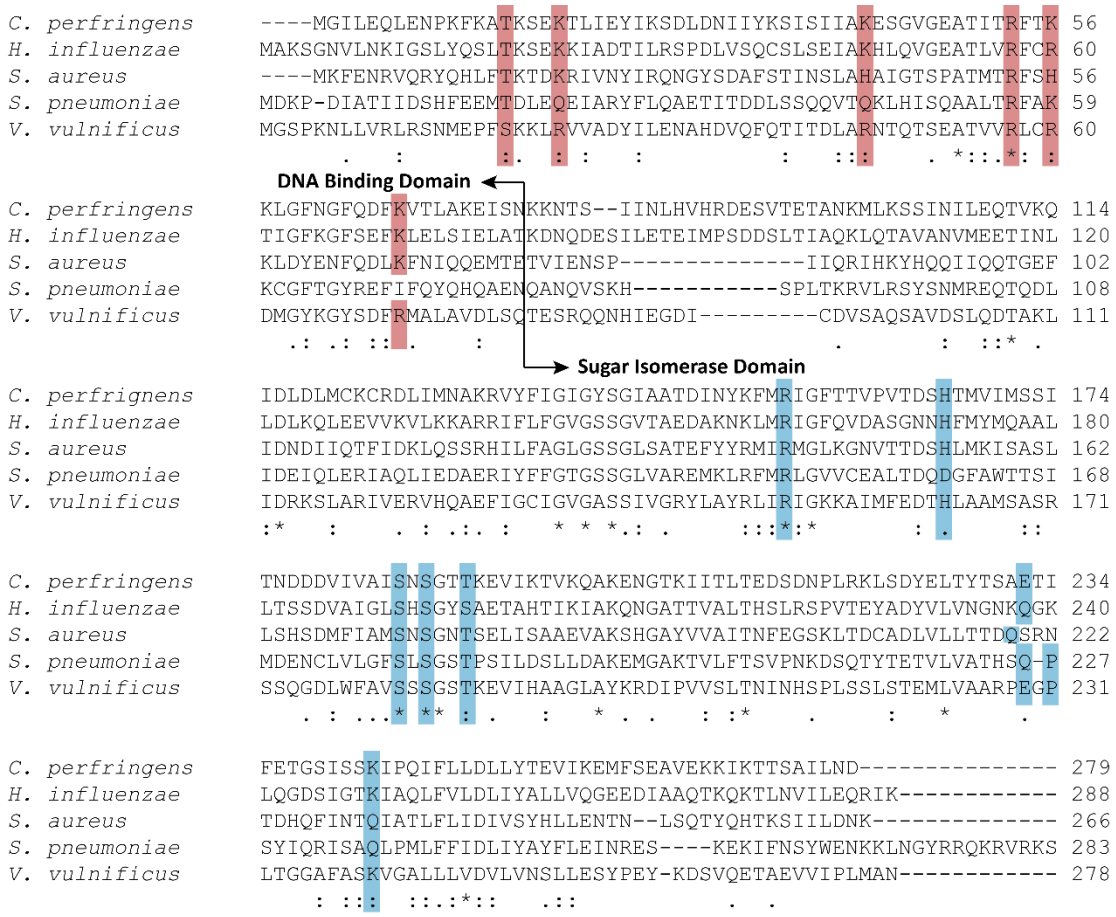


Figure 4.7 | Multiple sequence alignment of RpiR-type sialoregulators from five bacterial pathogens. Species include *C. perfringens*, *H. influenzae*, *S. aureus*, *S. pneumoniae* and *V. vulnificus*. Strictly conserved residues (*****), conservation between residues with strongly similar properties (**:**) and conservation between residues with weakly similar properties (**.**) are shown. Highlighted in red are residues that may be functionally important for DNA interaction. Highlighted in blue are residues that may be functionally important for binding the effector. These assignments are based upon the *V. vulnificus* structure (PDB ID - 4IVN). Labelled with black arrows is the N-terminal DNA-binding domain and the C-terminal sugar isomerase domain.

Overall, this multiple sequence alignment identified amino acid residues that are likely to coordinate the effector and highlighted their conservation with the *V. vulnificus* structure, providing further evidence that *N*-acetylmannosamine-6-phosphate is the effector molecule in *S. pneumoniae* NanR.

4.3.4 The quaternary structure of *S. pneumoniae* NanR

Sedimentation velocity experiments were carried out to characterise the quaternary structure of *S. pneumoniae* NanR. In order to extract this information, the purified protein was subjected to a centrifugal force field causing separation of the components within the sample based on their size and shape (34). This process can be monitored by absorbance optics to measure the sedimentation and diffusion behaviour of the macromolecules in solution. In addition, this solution behaviour can be evaluated across a wide range of concentrations and buffer conditions (35).

Sedimentation data was collected for NanR at a concentration of 0.25 and 0.5 mg mL^{-1} (Chapter Eight, Section 8.6.2.5) and fitted to a continuous sedimentation coefficient distribution [c(*s*)] model, resulting in a single, symmetrical peak. Peak integration within the c(*s*) distribution gave a sedimentation coefficient of 4.31 S for *S. pneumoniae* NanR (Figure 4.8, Panel A). The frictional ratio (f/f_0), which is a measure of shape asymmetry suggested that NanR (f/f_0 = 1.28) has a slightly elongated shape in solution, as a sphere would have a f/f_0 of 1.0.

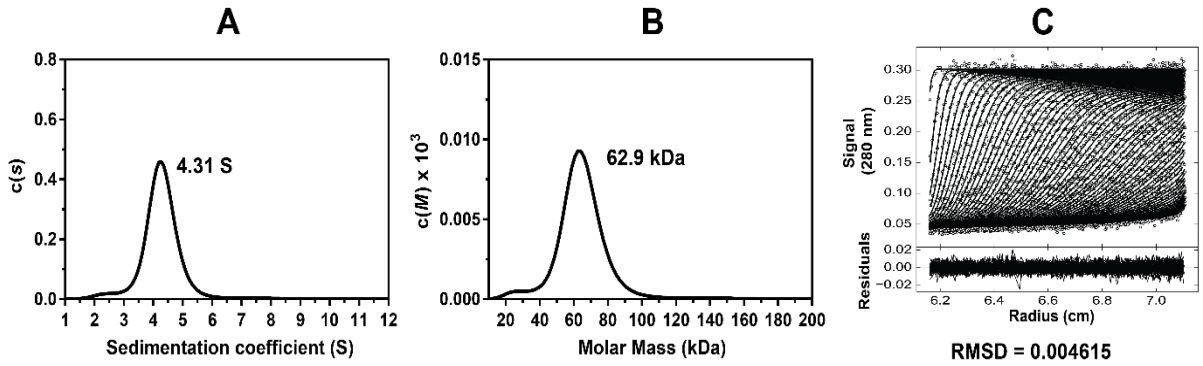


Figure 4.8 | Sedimentation velocity analysis of *S. pneumoniae* NanR using analytical ultracentrifugation. Data was collected at 280 nm, 42 000 rpm and 20 °C using purified NanR at a concentration of 0.5 mg mL^{-1}. **Panel A)** The c(*s*) model displaying the sedimentation distribution. Shown in bold is the sedimentation coefficient, following peak integration. **Panel B)** The c(*M*) model displaying the molar mass distribution. Shown in bold is the apparent molar mass, following peak integration. **Panel C)** The raw data (open circles) and best fit (lines) is presented. For clarity, every third absorbance scan is shown. The randomly distributed residuals are shown below. All data was analysed using SEDFIT (36). Figures were produced using PRISM (version 7.00) and GUSSI (developed by Dr Chad Brautigam).

Further fitting to a continuous mass distribution [c(*M*)] model gave an apparent molar mass of 62.9 kDa, consistent with the theoretical dimeric mass of 65.3 kDa (Figure 4.8, Panel B). The observation of a dimeric assembly is a common feature amongst bacterial transcriptional regulators (37; 38). Although a smaller species was observed in the sedimentation data at ~2.5 S, the ratio between this smaller peak and the predominant peak did not change with concentration. This is consistent with the conclusion

that NanR is not in a self-association and suggested that this peak is likely to be a contaminant or an incompetent monomer of NanR. However, it could be a very slow self-association.

Combined, these sedimentation velocity data confirmed that NanR adopts a dimeric architecture in solution, which is consistent with the reported oligomeric structure of *V. vulnificus* NanR (13). Moreover, the observation of a predominant, symmetrical peak demonstrated NanR is stable in solution. Validation of this analysis was supported by the randomly distributed residuals and a low root-mean-square deviation (Figure 4.8, Panel C).

4.3.5 Electrophoretic mobility shift assays demonstrate DNA binding activity

Electrophoretic mobility shift assays (EMSA) were conducted using a nanomolar titration range of protein against a set concentration of fluorescently end-labelled DNA as described in Chapter Eight, Section 8.6.2.7, in order to identity DNA-binding activity. This double-stranded oligonucleotide contains the DNA recognition sequence found within the operator site of the *S. pneumoniae nan* operon I (Figure 4.3) and is labelled with fluorescein at the 5' terminal. In addition, these experiments were replicated in the presence of *N*-acetylmannosamine-6-phosphate to further investigate the role of this phosphorylated sugar as the effector molecule for *S. pneumoniae* NanR.

Visualised using a fluorescent scanner, the addition of NanR resulted in a concentration-dependent gel shift, relative to the DNA probe. This single observed band, which increases in signal intensity is consistent with the formation of a single protein-DNA hetero-complex (Figure 4.9A). However, as the protein concentration was increased further, this band disappeared leaving a smear in the EMSA gel (Figure 4.9A, Lanes 9-11). Instead, the sample was observed at the bottom of the well suggesting that the species has significantly increased in size and is now unable to enter the EMSA gel. In addition, a reduced signal intensity of the DNA probe was observed in lanes 9-11, respectively (Figure 4.9A). Together, these features suggest the formation of non-specific DNA-binding within this titration range. It should be noted that the intensity of the retarded band for the hetero-complex was relatively low compared to the DNA probe. Further optimisation of the assay's conditions may enable this signal to be increased.

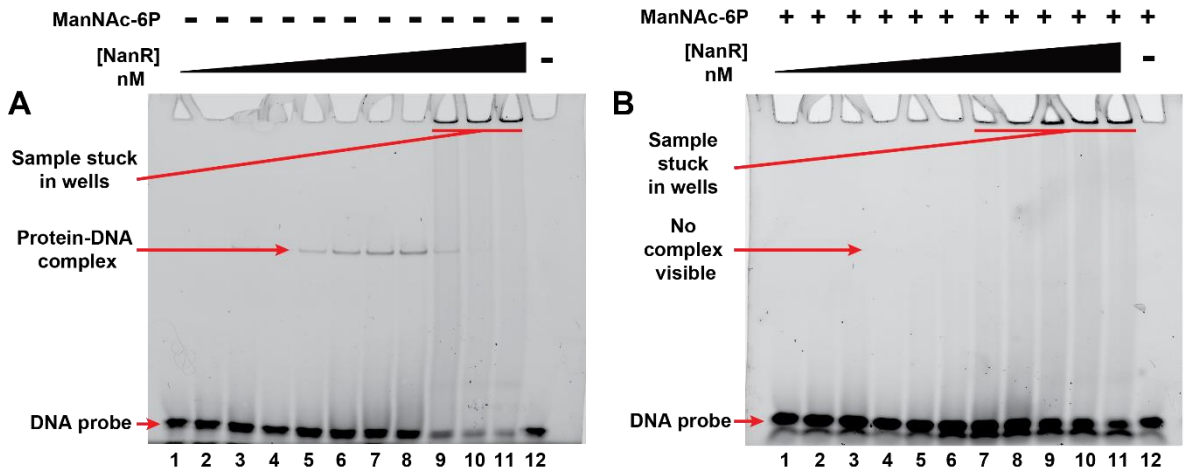


Figure 4.9 | Electrophoretic mobility shifty assays in the presence and absence of N-acetylmannosamine-6-phosphate. A) Titration range of NanR in the presence of 50 nM DNA. **B)** Titration range of NanR in the presence of 50 nM DNA and 5 mM *N*-acetylmannosamine-6-phosphate (ManNAc-6P). In both gels each lane contains as follows: Lane 1 – 12.5 nM NanR; Lane 2 – 25 nM NanR; Lane 3 – 50 nM NanR; Lane 4 – 60 nM NanR; Lane 5 – 70 nM NanR; Lane 6 – 80 nM NanR; Lane 7 – 90 nM NanR; Lane 8 – 100 nM NanR; Lane 9 – 200 nM NanR; Lane 10 – 300 nM NanR; Lane 11 – 400 nM NanR; and Lane 12 – DNA only. Labelled is the DNA probe, the protein-DNA hetero-complex and evidence of sample stuck in the wells.

Interestingly, when *N*-acetylmannosamine-6-phosphate was included in the binding assay, a complete loss of activity was observed (Figure 4.9B). This feature provides strong evidence to reinforce that this phosphorylated sugar is the effector molecule as its presence has significantly attenuated the DNA-binding affinity for NanR towards its recognition sequence. Although, relative to the previous experiment, more sample was observed at the bottom of the well, without a clear reduction in signal for the DNA probe (Figure 4.9B, Lanes 7-11). Intriguingly, this may reflect aggregated protein as opposed to non-specific binding to DNA as no clear retardation of DNA was present.

Nonetheless, these EMSA experiments demonstrate that NanR can actively bind DNA and illustrate that in the presence of *N*-acetylmannosamine-6-phosphate this DNA binding activity is lost or significantly attenuated. To further unravel the molecular determinants for this attenuation, a crystallographic analysis in the presence and absence of *N*-acetylmannosamine-6-phosphate is required. This would enable any effector-induced conformational changes to be mapped, in order to understand the molecular choreography that occurs as part of the allosteric regulation in *S. pneumoniae*.

4.3.6 Defining the stoichiometry of the *S. pneumoniae* NanR-DNA hetero-complex

The stoichiometry of protein-DNA hetero-complexes is crucial to developing a molecular understanding of the DNA-binding mechanism. As mentioned earlier, sedimentation velocity experiments conducted using an analytical ultracentrifugation provide access to the sedimentation and diffusion behaviour of macromolecules in solution (35). In addition, analytical ultracentrifugation has long been recognised as the gold-standard technique to quantitatively analyse complex interactions in solution, since experiments can be performed in conditions that closely match the physiological environment (39; 40). Thus, sedimentation velocity experiments can provide a qualitative assessment of EMSA assays, which can be limited in terms of measuring the dissociation of the complex due to a non-equilibrium environment. Furthermore, EMSA assays do not provide accurate determination of the molecular mass for the complexes, as the gel-based technique is influenced by several other factors (41).

To investigate the nature of the *S. pneumoniae* NanR-DNA interaction, sedimentation velocity experiments were conducted using different titrations of the sialoregulator against a set concentration of fluorescently end-labelled DNA (Chapter Eight, Section 8.6.2.6) This double-stranded oligonucleotide is identical to what was in the EMSA experiments above. By measuring the emission wavelength of fluorescein (495 nm) the sedimentation profile of DNA can be monitored as each titration is subjected to a centrifugal force field. Thus, this not only enables detection of free DNA but also provides the ability to evaluate the formation of protein-DNA hetero-complex, indicated by a positive shift in the sedimentation coefficient.

To accurately determine the molecular weight of each species in solution and therefore gain access to the stoichiometry of this interacting system, three key pieces of information must be obtained, including: 1) sedimentation coefficient; 2) diffusion coefficient; and 3) partial specific volume. To precisely acquire the sedimentation and diffusion coefficients, sedimentation velocity experiments can be collected at two different speeds. A high speed enables the macromolecules within a complex mixture (such as protein and DNA) to be separated with increased resolution, resulting in an accurate sedimentation coefficient. Conversely, a lower speed allows more time for macromolecules to diffuse in solution, providing a more precise diffusion coefficient. However, each speed must be optimised for each experiment. This is dependent on the number of samples being collected and the approximate molar mass range of the macromolecules in solution. In this study, both 50 000 rpm and 32 000 rpm were selected. Lastly, the partial specific volume of the protein-DNA hetero-complex was estimated as a weight-average of the protein and DNA components (Chapter Eight, Equation 8.4). The values used in this study are presented in Table 4.2.

Following data collection, the sedimentation data was evaluated by the two-dimensional spectrum analysis (2DSA) (42) and further refined by 50 Monte Carlo iterations using the state-of-the-art software, Ultrascan (43). Using peak integration, the hydrodynamic information was extracted from both the high and low speed experiments and was used to define the stoichiometry. This revealed that upon increasing protein concentration, *S. pneumoniae* NanR coordinates DNA as a dimer with a 2:1 binding stoichiometry to form a dimeric NanR-DNA hetero-complex (Figure 4.10).

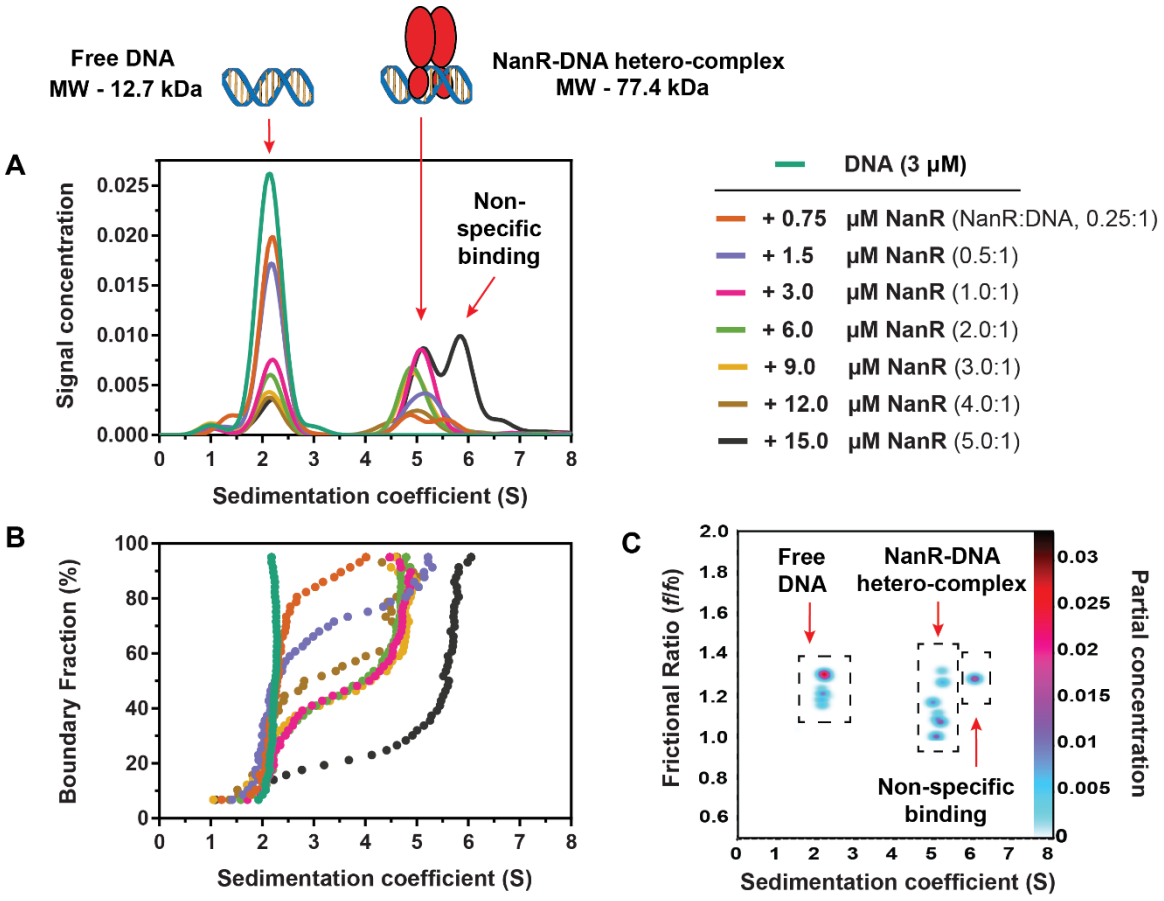


Figure 4.10 | Single-wavelength sedimentation velocity analysis of NanR from *S. pneumoniae*. A) 2DSA-Monte Carlo model displaying the sedimentation coefficient (S) distribution from the 50 000 rpm experiment, where each colour represents a different titration sample, collected independently (see legend). To illustrate the identity of the two main species in solution, a cartoon depiction of DNA and the NanR-DNA hetero-complex is presented. **B)** van Holde-Weischet plot provides a graphical transformation of the sedimentation data allowing the proportional contributions of each species to be interpreted. **C)** Pseudo-3D plot presenting the global solute distribution, across all sedimentation data shown in (A) as a function of frictional ratio and sedimentation coefficient.

Across the seven different NanR-DNA titrations measured in this experiment, all displayed a positive shift in the sedimentation coefficient, relative to the DNA signal at 2.14 S (Figure 4.10A). This shift can clearly be observed in the van Holde-Weischet plot (Figure 4.10B), which provides a graphical transformation of the sedimentation data allowing the proportional contributions of each species to be interpreted. Moreover, upon increasing NanR concentration a new peak at approximately 5 S appears and a decrease in the relative DNA signal was observed—this is indicative of protein-DNA hetero-complex formation. The measured molar mass of this complex (reported in Table 4.2) was consistent with the theoretical molar mass of a NanR dimer and its DNA recognition sequence forming a dimeric NanR-DNA hetero-complex (77.4 kDa). Validation of this stoichiometry was supported by the small difference in molar mass between these values with an error of less than 10% (Table 4.2).

Table 4.2 | Summary of sedimentation velocity data collection and analysis from NanR *S. pneumoniae*

Data collection parameters						
Buffer density (g/cm^3)					1.0059	
Buffer viscosity (cp)					1.0211	
Partial specific volume (mL/g)						
S. pneumoniae NanR					0.7354	
DNA oligonucleotide					0.5550	
Dimeric NanR-DNA hetero-complex					0.7065	
Dimeric NanR-DNA hetero-complex (+ additional DNA)					0.6854	
Sample	**Sedimentation Coefficient (x10^{-13} S)**	**Diffusion Coefficient (x10^{-07} D)**	**Measured Molar mass (kDa)**	**Oligomeric state of hetero-complex**	**Expected Molar mass (kDa)**	**% Error (Molar mass)**
DNA only	2.14	10.35	12.5	-	12.1	3.3
0.25:1	5.19	5.55	77.3	Dimeric	77.4	0.1
0.5:1	5.15	5.44	78.3	Dimeric	77.4	0.5
1:1	5.08	5.33	78.8	Dimeric	77.4	1.8
2:1	4.93	5.47	74.5	Dimeric	77.4	3.7
3:1	4.97	5.71	72.0	Dimeric	77.4	6.9
4:1	5.07	5.97	70.1	Dimeric	77.4	9.4
5:1	5.53	4.71	90.5	Dimeric + DNA	89.5	1.1

However, at a protein-DNA ratio of 5:1, a new peak was observed with a small positive shift in the sedimentation coefficient. Following further analysis, the measured molar mass of this species was consistent with the theoretical molar mass of the dimeric NanR-DNA hetero-complex with an additional DNA oligonucleotide, as opposed to a second NanR dimer. This observation suggests that non-specific binding may occur between neighbouring DNA oligonucleotides at high concentration, or upon saturating conditions of the hetero-complex. A similar observation at this protein-DNA molar ratio was observed in the EMSA experiment, where the hetero-complex did not enter the gel and remained at the bottom of the well (Figure 4.9A, Lanes 9-11). Further optimisation, with respect to ionic strength or pH of the buffer system used in the experiment may reduce the appearance of the non-specific binding.

The global distribution of the components in this experiment can be visualised using a Pseudo-3D plot (Figure 4.10C). Here, a spot is representative of a solute, where upon increasing concentration the intensity of that spot changes. For this experiment, three populations were visualised: 1) free DNA; 2) NanR-DNA complex; and 3) non-specific binding. While the frictional ratio was seen to vary for the protein-DNA hetero-complex, this was likely due to the presence of a reaction boundary during complex formation.

To determine the binding kinetics, the ratio of bound and unbound DNA was determined by peak integration within the sedimentation coefficient distribution. The data was best fit to a single-site model in Prism (Figure 4.11) and gave a dissociation constant of 0.91 µM (95% confidence interval of 0.64, 1.29 µM) for the NanR-DNA interaction. This demonstrated that NanR has a low micromolar affinity for its DNA recognition sequence.

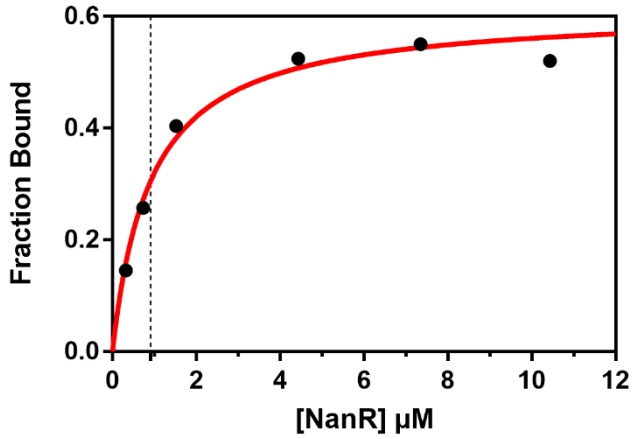


Figure 4.11 | The DNA binding isotherm for *S. pneumoniae* NanR. The binding isotherm from the sedimentation data, best fit to a single-site model provides a dissociation constant (K_D) of 0.91 µM (black dash).

Taken together, these sedimentation velocity experiments have defined the stoichiometry of NanR-DNA hetero-complex formation in *S. pneumoniae*, supported by the small difference between measured and theoretical molar mass values, and have determined that this system displays high affinity for its DNA recognition sequence. Overall, this demonstrated the precision of analytical ultracentrifugation to accurately provide hydrodynamic information in order to characterise macromolecules in solution and reinforces why it is regarded as the gold-standard technique to quantitatively analyse complex interactions in solution.

4.3.7 Crystallisation studies of *S. pneumoniae* NanR

To gain insight into the molecular mechanism of gene regulation, *S. pneumoniae* NanR was subjected to crystallisation studies, following purification, with the aim of obtaining a crystal structure. Here, crystallisation trials were conducted using the sitting-drop vapor-diffusion method, as described in Chapter Eight, Section 8.6.2.8, across multiple crystallisation screens from Molecular Dimensions, including JCSG-*plus*, PACT-*premier*, Shotgun, Morpheus and the Clear Strategy Screens (I and II). In addition, the effector molecule, *N*-acetylmannosamine-6-phosphate (Section 4.3.3) was included in the protein-buffer solution and subjected to co-crystallisation trials in parallel. Here, the aim was to gain functional insight into coordination of this phosphorylated sugar. Collectively, these crystallisation screens were incubated at two temperatures (8 °C and 20 °C) and were frequently monitored for any signs of crystal growth.

Over several weeks, various wells across the crystallisation trials presented the appearance of precipitation, phase separation and spherulites, while very few showed signs of promising crystal development. However, following a further incubation, crystals that were suitable for diffraction screening were identified at 20 °C. Collectively, these crystals were obtained from very different conditions, where one grew in Shotgun condition C6 (2.0 M ammonium sulphate, 0.1 M sodium acetate, pH 4.6), while the other developed in PACT-*premier* condition C9 (0.2 M lithium chloride, 0.1 M HEPES, pH 7.0, 20 % (w/v) PEG 6000). In both cases, the protein-buffer solution contained 10 mM *N*-acetylmannosamine-6-phosphate. No crystal growth was seen in conditions without this phosphorylated sugar.

Figure 4.12A and 4.12B show the difference in crystal morphology between Shotgun condition C6 and PACT-*premier* condition C9, respectively. When observed under a polarising filter, both crystal morphologies were birefringent, suggesting each crystal was protein. Excitingly, when collected using the MX2 beamline at the Australian Synchrotron, these crystals diffracted to 3.30 Å and 3.72 Å, respectively. The statistics for each dataset following data processing are summarised in Table 4.3.

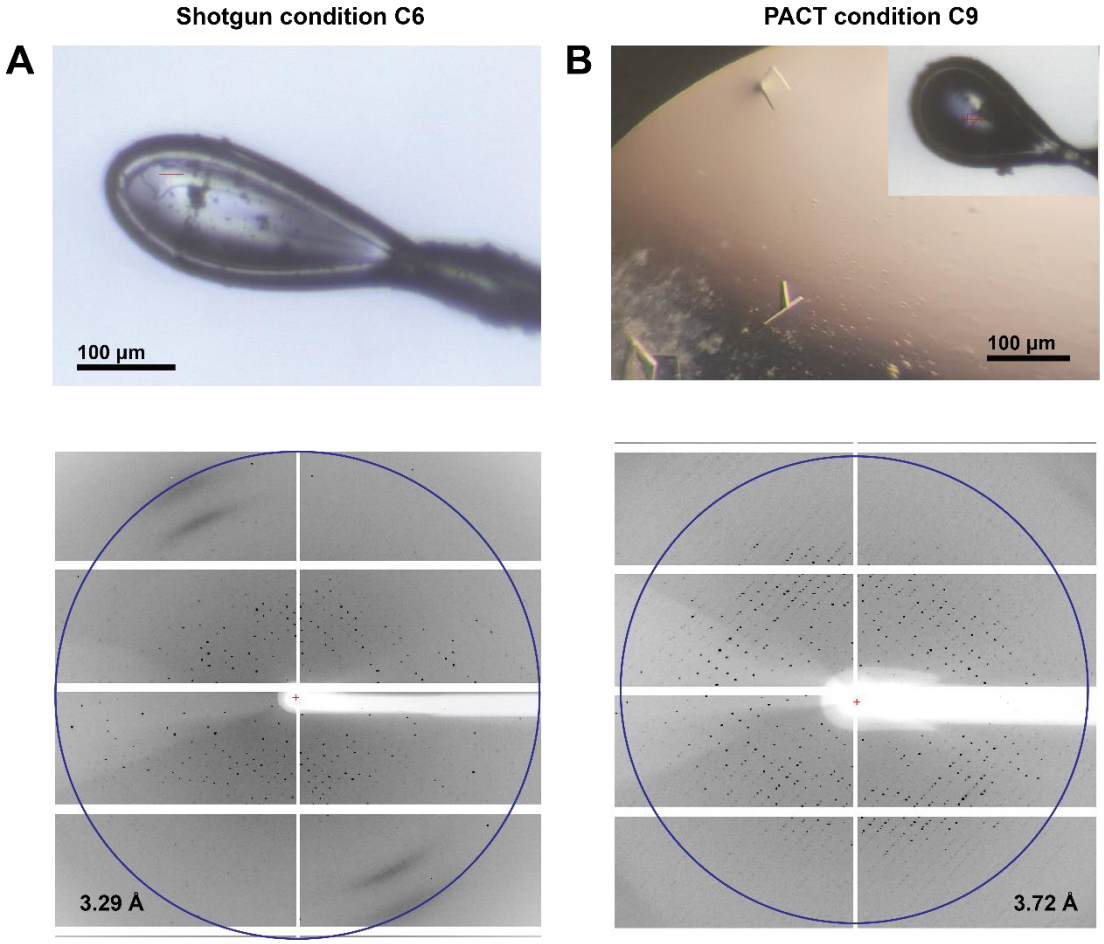


Figure 4.12 | Crystal images and X-ray diffraction images of *S. pneumoniae* NanR. A) Crystal obtained from Shotgun condition C6 that diffracted to 3.29 Å. **B)** Crystal obtained from PACT-*premier* condition C9 that diffracted to 3.72 Å. The red crosshair highlights the crystal mounted within a cryo-loop on the MX2 beamline. A scale bar is included to give an indication on the size of each crystal. The blue ring indicates where the highest diffraction was observed in each crystal.

Table 4.3 | Data collection statistics from *S. pneumoniae* NanR crystallisation trials*

	Shotgun condition C6	PACT-*premier* condition C9
Data collection statistics		
Spacegroup	H32	P21
Unit cell *a*, *b*, c (Å)	124.94, 124.94, 193.62	126.58, 218.58, 127.55
Unit cell α, β, γ (°)	90, 90, 120	90, 93.68, 90
Resolution range (Å)	47.23-3.55 (3.29-3.29)	49.43-3.80 (3.72-3.72)
No. of total reflections	46, 285 (8, 984)	249, 764 (13, 822)
No. of unique reflections	8, 889 (1, 817)	71, 405 (3, 950)
I/σ	7.6 (1.8)	3.2 (0.6)
Completeness (%)	98.4 (99.6)	98.4 (85.4)
$CC_{1/2}$ (%)	99.3 (70.2)	77.6 (9.0)
R_{merge}	0.154 (0.709)	0.689 (5.606)
R_{pim}	0.078 (0.400)	0.509 (4.400)

*Values in parentheses are for the highest resolution shell

Unfortunately, both these datasets were not suitable for structure determination of *S. pneumoniae* NanR for several reasons. Firstly, both were susceptible to radiation damage during data collection. This was significant in the PACT-*premier* condition C9, indicated by the poor correlation coefficient ($CC_{1/2}$) and high R_{merge} value. Moreover, these crystals were very thin on one axis and hence were difficult to centre, which contributed to the reduced level of completeness and poor I/σ value. Secondly, molecular replacement models for these proteins are limited as there are only three RpiR-like transcriptional regulators in the PDB, which all have a sequence identity less than 21%. This low identity is largely due to the sugar isomerase domain as suggested in Figure 4.7. When combined with a lower resolution dataset, this makes phasing the data increasingly difficult. Further, neither an ensemble-based molecular replacement strategy, nor an automated crystal structure determination platform, such as Auto-Rickshaw (44) or MrBUMP (45), were successful at phasing the diffraction data.

However, these two conditions are perfectly suited to optimisation in the future. This can be achieved by screening the salt, buffer, or precipitant around each condition. Moreover, crystal seeding would provide an initial nucleation event to encourage larger crystal growth, with the aim of reducing sensitivity to radiation damage and simplifying the alignment of crystals. Lastly, the protein could be labelled with selenomethionine to provide an alternative phasing strategy by single-wavelength anomalous diffraction (46). Most importantly, obtaining this diffraction data from an initial crystallisation screen, combined with the level of resolution observed, demonstrates that optimisation and ultimately structure determination is feasible for this protein.

4.3.8 Small angle X-ray scattering of *S. pneumoniae* NanR

In the absence of an atomic resolution structure, I employed an alternative strategy, small-angle X-ray scattering (SAXS) to gain a structural understanding of the RpiR-type sialoregulators. Although low-resolution (10-50 Å), SAXS can provide insight into the size and shape of biological macromolecules in solution and enable measurement of the molar mass and the molecular envelope of proteins, DNA or complexes (47; 48). Together, this information can be used to complement hydrodynamic information from analytical ultracentrifugation or alternatively the crystal packing within an atomic resolution structure to provide confirmation of the biologically relevant quaternary structure (48; 49). In addition, SAXS can provide a gauge on the flexibility or disorder of biological macromolecules and allow interpretation of conformational changes that may occur upon addition of a ligand (such as DNA) (49).

Since the total solution scattering is a feature of all species within the sample, including the buffer the presence of aggregates or contaminants can have a significant effect on the overall quality (50). Thus, to help mitigate these effects, all samples were subjected to in-line size-exclusion chromatography prior

to SAXS data collection using the SAXS/WAXS beamline at the Australian Synchrotron as described in Chapter Eight, Section 8.6.2.9. Here, the solution structure of *S. pneumoniae* NanR was determined (Figure 4.13). In addition, SAXS experiments were conducted in the presence of DNA (Figure 4.14), in order to investigate any conformational changes and visualise the hetero-complex through multiphase modelling (Figure 4.16). Together, these solution-scattering experiments are discussed below.

4.3.8.1 The solution structure of S. pneumoniae NanR

SAXS data collected for the *S. pneumoniae* sialoregulator was consistent with a dimeric assembly in solution. This supports the sedimentation velocity data obtained in Section 4.3.5, verifying the quaternary structure. The scattering data is presented below in Figure 4.13, while a summary of the data analysis is provided in Table 4.4. Guinier analysis produced a linear plot (Figure 4.13A, insert) with a radius of gyration (R_g) of 34.56 Å and a forward scattering value (I_0) of 0.024 (Table 4.4). The linearity of the Guinier plot further supports the observation of a single monodisperse species in solution, free of any significant amount of aggregation or inter-particle interference. The maximum inter-particle dimension (D_{max}) was calculated as 112 Å by an in-direct Fourier transform of the scattering data (Figure 4.13C). Kratky analysis presented a bell-shaped curve (Figure 4.13D), which indicated the protein was folded in solution and provided evidence of some flexibility. The molecular weight determined *via* the Porod volume was 77.6 kDa, which was consistent with a dimeric assembly in solution, albeit slightly larger than the theoretical dimeric mass (65.3 kDa). However, the molecular weight determined used the online server SAXS-MoW2 (51) was 91.2 kDa, which was closer to the theoretical trimeric mass (98.0 kDa). As there was no sign of aggregation in the Guinier plot or indication of a trimeric species during sedimentation velocity experiments (Section 4.3.5, 4.3.7), this discrepancy in molecular weight is likely attributable to several features: 1) the low resolution of SAXS data (52); 2) the flexibility observed in NanR, as supported by the Kratky analysis; and 3) the elongated shape of NanR in solution. This is supported by the frictional ratio of 1.28, calculated during sedimentation velocity experiments (Section 4.3.5) and the positively-skewed tail of the pairwise distribution plot (Figure 4.13C).

Table 4.4 | Summary of SAXS data analysis for *S. pneumoniae* sialoregulator, DNA and the hetero-complex

SAXS data analysis	S. *pneumoniae* NanR	DNA only	NanR-DNA Complex
Guinier analysis			
I(0) (cm-1)	0.0240 ± 0.0004	0.0030 ± 0.00002	0.0260 ± 0.0001
R_g (Å)	34.56 ± 0.94	17.44 ± 1.33	38.78 ± 0.71
P(*r*) analysis			
I(0) (cm-1)	0.024 ± 0.002	0.0030 ± 0.00004	0.0260 ± 0.0002
R_g (Å)	34.75 ± 0.22	17.21 ± 0.15	39.45 ± 0.41
D_{max} (Å)	112	50	133
Porod volume ($Å^{-3}$)	132 000	11 900	127 000
Molecular Mass (MM) estimation, kDa			
MM SAXS-MoW2[a]	91.2	9.8	90.3
MM Porod volume	77.6	7.0	74.7

[a] http://saxs.ifsc.usp.br/ (51)

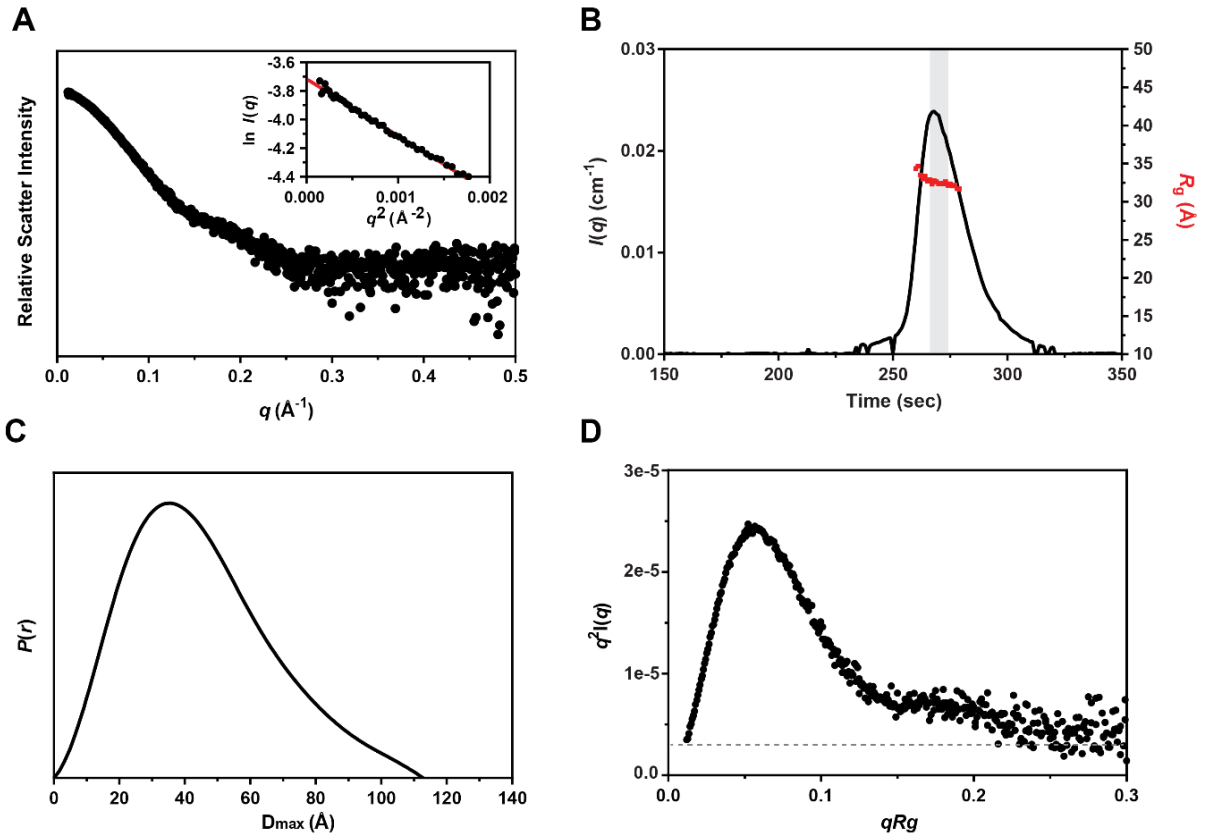


Figure 4.13 | SAXS data for the *S. pneumoniae* sialoregulator. A) Solution scattering data presented as an intensity plot. The Guinier plot shown as an insert, highlights linearity at low *q*—this provides confidence that the data is free of any significant amount of aggregation or inter-particle interference. **B)** Plot showing the forward scattering value *I*(0) (black line) and radius of gyration (*Rg*) (red squares) as a function of time. The grey box highlights the frames that were selected for averaging to produce the plot shown in A. **C)** Pairwise distribution plot displaying the maximum inter-particle dimension (D_{max}) as 112 Å. **D)** The Kratky plot presents a bell-shaped curve, which indicates protein is folded and provides evidence of some flexibility in the structure, due to deviation from the baseline (grey dash) between 0.1 and 0.2 *q*.

4.3.8.2 The solution structure of the S. pneumoniae NanR-DNA hetero-complex

Having characterised the solution structure of the *S. pneumoniae* sialoregulator and defined the stoichiometry of the NanR-DNA hetero-complex through analytical ultracentrifugation, further SAXS experiments were conducted in the presence of DNA in order to characterise the solution structure of this hetero-complex. Here, protein and DNA were mixed at a molar ratio of 2:1 and injected onto the in-line size-exclusion chromatography column prior to data collection. This allowed separation of the NanR-DNA hetero-complex from the interaction partners, any non-specific binding, which was observed in the EMSA experiments (section 4.3.6) and the presence of any aggregation within the sample, while ensuring the sample was buffer matched. To illustrate this process, Figure 4.14 presents the UV trace

from the in-line size exclusion column and associated forward scattering intensity/ *Rg* plot for NanR, DNA and the hetero-complex. Here, a positive shift to the left was observed in the presence of DNA compared with NanR by itself, indicating a larger species. Moreover, this shifted peak had a higher absorbance at 260 nm (beige line) than 280 nm (red line). Taken together, this demonstrates protein-DNA hetero-complex formation. The solution scattering data for the NanR-DNA hetero-complex and the DNA interaction partner is presented in Figure 4.15, while all data analysis is summarised in Table 4.4.

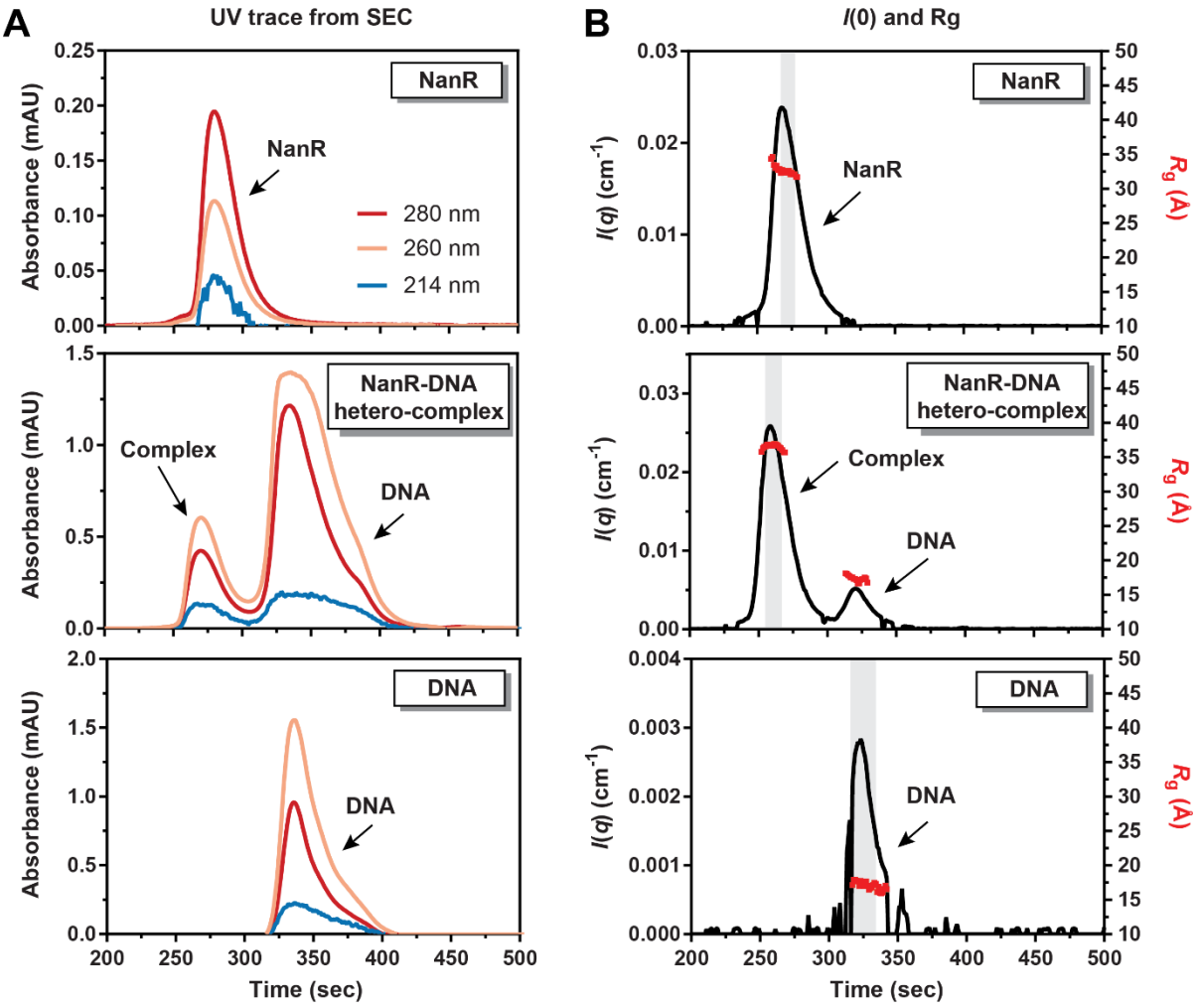


Figure 4.14 | SAXS data of *S. pneumoniae* NanR, the NanR-DNA complex, and DNA from the in-line size exclusion column. A) UV trace from the in-line size exclusion column. The absorbance at 280 nm (red line), 260 nm (beige line) and 214 nm (blue line) is shown. **B)** Plot showing the forward scattering value *I*(0) (black line) and radius of gyration (*R*g) (red squares) as a function of time. The grey box highlights the frames that were selected for averaging to produce scattering curve shown in Figure 4.13A.

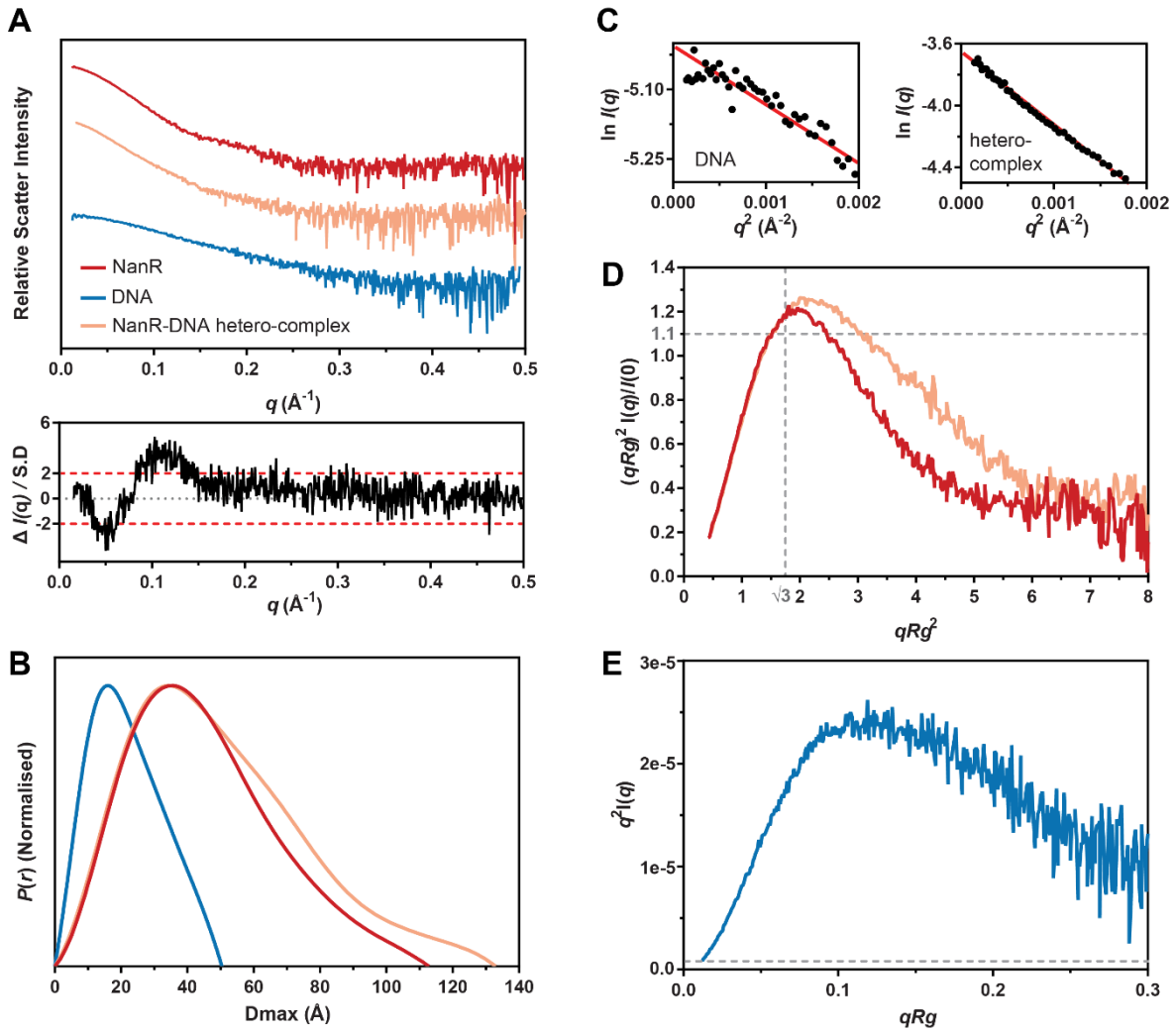


Figure 4.15 | SAXS data for the *S. pneumoniae* sialoregulator in the presence of DNA. A) Solution scattering data presented as an intensity plot for NanR only (red), DNA only (blue) and the NanR-DNA hetero-complex (beige). Shown below is a residuals plot to highlight the difference in solution scatter between the protein only (red) and complex (orange). This difference between the experimental $I(q)$ values is divided by the associated experimental standard deviation (S.D.) and plotted as a function of q. The two horizontal red lines indicate ± 2 S.D., which is used as limits for a good fit within the program US-SOMO (53). **B)** Pairwise distribution plot displaying the difference in maximum inter-particle dimension (D_{max}) between the complex and each interacting partner. **C)** The Guinier plot for the DNA and hetero-complex data is shown as an insert, highlights linearity at low q—this provides confidence that the data is free of any significant amount of aggregation or inter-particle interference. **D)** Dimensionless Kratky plot suggests that the complex has a slightly more elongated shape than NanR, as the peak maximum has shifted to further to the right. A typical globular protein has a peak maximum at 1.1 as highlighted by the grey dashes. **E)** The Kratky plot for the DNA highlights the inherent flexibility of the oligonucleotide in solution having lost the bell-shaped curve of a globular particle and not having convergence to the baseline (grey dash). While the scattering data of NanR is the same as Figure 4.13, it is presented here for clarity.

A comparison between the SAXS data of NanR and the NanR-DNA hetero-complex demonstrated that in the presence of DNA, NanR undergoes a conformational change in solution. This was indicated by several features including: 1) subtle changes to the solution scatter through the mid *q* range (0.1-0.2); 2) an increase in *Rg* from 34.56 Å to 38.78 Å; and 3) a more elongated particle in solution, suggested by the larger inter-particle dimension (D_{max}) of 133 Å, as opposed to 112 Å (Figure 4.15B), and the right-shifted peak maximum in the dimensionless Kratky plot (Figure 4.15D).

The molecular weight determined *via* the Porod volume was 74.7 kDa, which was consistent with a theoretical dimeric NanR-DNA hetero-complex mass (77.4 kDa). In addition, the molecular weight determined by the online server SAXS-MoW2 (51) was 90.3 kDa, which was also consistent with a dimeric NanR-DNA hetero-complex in solution, albeit slightly larger than the theoretical mass .

Conversely, DNA was much smaller with a molecular weight determined by the online server SAXS-MoW2 (9.8 kDa) and the Porod volume (7 kDa) that was consistent with the theoretical mass (12.1 kDa) for the double-stranded oligonucleotide. In solution, this oligonucleotide displayed a significant amount of flexibility, as suggested by the large rise and lack of a bell-shaped curve in the Kratky plot (Figure 4.15E). This observation is understandable, as in solution the oligonucleotide would sample multiple different conformations and twists of the double helix. While the UV absorbance at 260 nm for the DNA sample was relatively high compared with the other samples, due to the aromaticity of the nucleic acid base pairs, its respective forward scattering intensity was much lower (Figure 4.14A and B, respectively). This is because the oligonucleotide has a much smaller molar mass and partial specific volume compared with NanR, or the hetero-complex and hence a smaller scattering length, leading to a reduced scattering amplitude.

4.3.8.3 Multiphase ab initio model of the S. pneumoniae NanR-DNA hetero-complex

To provide further structural insight into the NanR-DNA hetero-complex, I reconstructed a 3D *ab initio* model (Figure 4.16) from the scattering curves of the complex, along with the interaction partners, protein and DNA using the multiphase modelling software MONSA (54). This process uses densely packed beads and multiple rounds of simulated annealing to find the best fit between the experimental scattering and the theoretical scattering of the *ab initio* model being built. Although, as these 3D *ab initio* models are derived from a 1D scattering profile they are inherently ambiguous, meaning models of very similar conformations can all be generated from the same experimental data, thus one model alone is not expected to provide a realistic structure at this resolution (54). This ambiguity can be reduced by reconstructing an ensemble of models and then averaging these using DAMAVER (55) to find the model that is the best representation of the scattering data. Moreover, each model generated

from multiple runs can be compared with each other using SUPCOMB to determine the level of consistency. The mean normalised spatial discrepancy (NSD) for reconstructions of the *S. pneumoniae* NanR-DNA hetero-complex was 0.70 ± 0.03. This low value demonstrates a good level of consistency between each independent model. As a result, the averaged reconstruction presented in Figure 4.16B provides qualitative evidence on the relative orientation of the protein and DNA when bound to each other in solution.

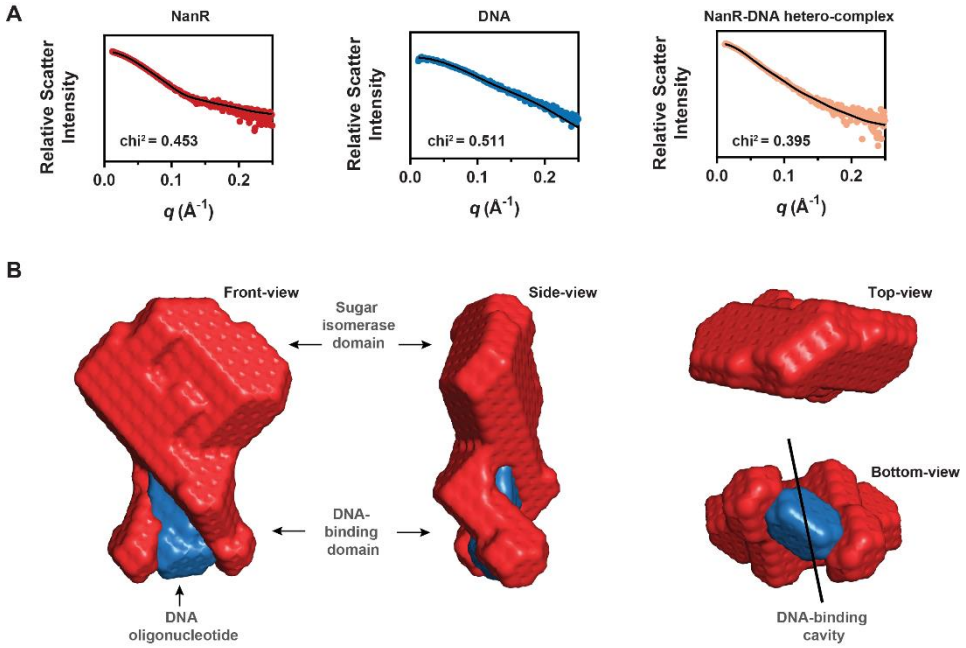


Figure 4.16 | Multiphase *ab initio* model of the *S. pneumoniae* NanR-DNA hetero-complex. A) The scattering profiles of the protein (red spheres), DNA (blue spheres) and protein-DNA complex (beige spheres) with MONSA generated fits (black line). The respective chi^2 value for each fit is shown **B)** The multiphase ab *initio* model of the NanR-DNA hetero-complex. Highlighted is the predicted sugar isomerase (SIS) domain, which is known to bind the effector molecule and the DNA-binding domain. Here, the DNA (in blue) appears to bind within a cavity formed by these DNA-binding domains (in red).

This *ab initio* model suggests that NanR binds DNA between two 'arms'. These are likely the DNA-binding domains from each monomer of the dimeric assembly, which interact with opposite sides of the DNA in order to accommodate the double helix within a binding cavity. The exact conformation of the DNA does seem implausible, as it would make sense for it to bind perpendicular rather than parallel to the protein, in order to accommodate the genomic DNA within the cell. However, this model is a low-resolution reconstruction and requires further structural evidence at higher resolution to verify the correct binding conformation. This can be explored in future experiments through X-ray crystallography or through cryo-electron microscopy in the presence of DNA. Nonetheless, this low-resolution model still provides qualitative insight into the relative orientation of DNA and binding between the DNA-binding domains, which is consistent with the function of the helix-turn-helix motif within this domain.

Combined, these SAXS experiments define the solution structure of *S. pneumoniae* NanR and provide insight into the conformational changes that occur upon DNA binding. While experiments in the presence of an effector molecule would help deduce any conformational changes that occur upon effector binding, these were not performed as making enough effector-supplemented buffer was not feasible for the in-line size exclusion SAXS experiments.

4.3.8.4 Comparison with Vibrio vulnificus NanR structure

To investigate if the solution structure of *S. pneumoniae* NanR is similar to the reported dimer of *V. vulnificus* NanR (13), a homology model was initially constructed for *S. pneumoniae* NanR by using the crystal structure of *V. vulnificus* NanR (PDB ID – 4IVN) as a template. CRYSOL (56) was subsequently used to generate a theoretical scattering profile of the homology model and then fit this to the experimental scattering profile of *S. pneumoniae* NanR in the presence and absence of DNA.

The architecture observed in the crystal lattice of *V. vulnificus* fit the experimental scatter of *S. pneumoniae* NanR with a chi^2 value of 2.0 (Figure 4.17A). This value supports that *V. vulnificus* is surprisingly similar to *S. pneumoniae* NanR in terms of oligomeric state, consistent with a dimeric architecture. In addition, this comparison suggests that *S. pneumoniae* NanR adopts a similar extended conformation in solution to *V. vulnificus* NanR, when in the absence of DNA. This is where the N-terminal DNA binding domains orient in opposing directions, while the C-terminal sugar isomerase domain serves as the oligomeric interface (Figure 4.17A, insert).

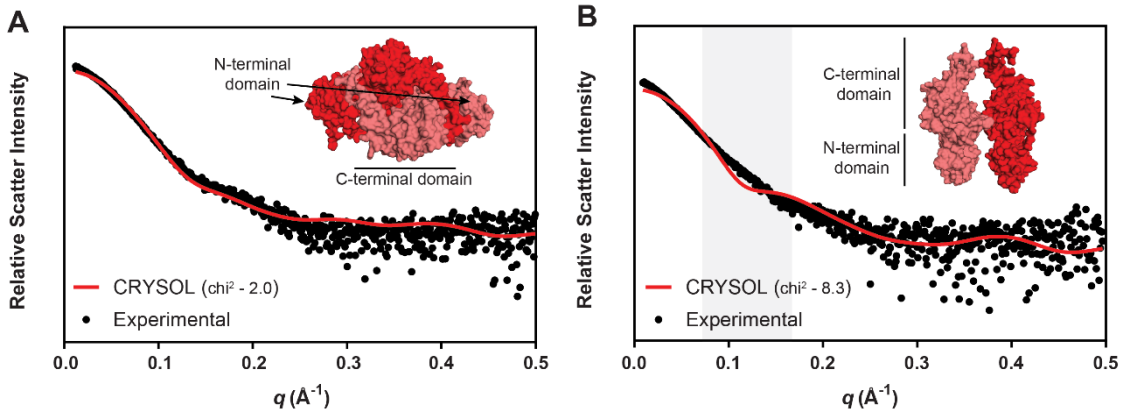


Figure 4.17 | Comparison of the solution structure for the *S. pneumoniae* and *V. vulnificus* sialoregulators. A) Small-angle X-ray scattering data of *S. pneumoniae* NanR (black sphere) is consistent with the theoretical scattering profile (red line, chi^2 value of 2.0) of the *V. vulnificus* crystal lattice architecture (insert). B) Small-angle X-ray scattering data of the *S. pneumoniae* NanR-DNA hetero-complex (black sphere) is inconsistent with the theoretical scattering profile (red line, chi^2 value of 8.3) of the *V. vulnificus* proposed 'active form' (insert). The *q* range with the main difference in the fit is highlighted (grey box). Each dimeric model is depicted as a surface representation.

This fit was repeated using a homology model of the authors proposed 'active form' of *V. vulnificus* NanR (13) using the experimental scatter of *S. pneumoniae* NanR-DNA hetero-complex, providing a chi^2 value of 8.3 (Figure 4.17B). In this model, the N-terminal DNA-binding domains are now oriented much closer together to create a DNA-binding cavity (Figure 4.17B, insert). However, the fit of this homology model suggested that *S. pneumoniae* NanR adopts an alternative shape in solution owing to differences in the mid *q* range (57). Despite this difference, both *S. pneumoniae* NanR and *V. vulnificus* NanR were largely consistent with a dimeric architecture in solution.

Overall, this comparison supports that *S. pneumoniae* NanR adopts a very similar conformation in solution to the crystal structure of *V. vulnificus* NanR, when in the absence of DNA. However, the poor fit of the proposed 'active form' of *V. vulnificus* NanR to the scattering profile of the *S. pneumoniae* NanR-DNA hetero-complex provides evidence to support that the model of *V. vulnificus* NanR is incorrect. Importantly, this evidence reinforces that the dimeric interface of the 'active form', which was interpreted from negative stain transmission electron microscopy data is unconvincing. To understand this further, it is important to note that the crystal structure of *V. vulnificus* NanR did not contain sufficient electron density to resolve the linker region between the N- and C-terminal domains, thus these domains are disconnected in the final model (PDB ID - 4IVN). This lack of density indicates this region is highly flexible and suggests it may play a role in facilitating the large conformational change of the N-terminal domain that is required to accommodate DNA within a binding cavity.

4.4 Chapter summary and model mechanism of regulation for *S. pneumoniae* NanR

This chapter developed a structural and functional understanding of the RpiR-type NanR from *S. pneumoniae*—the gene regulator of bacterial sialic acid catabolism.

Taken together, the stoichiometry, binding kinetics and solution structure for *S. pneumoniae* NanR, lead to a model mechanism for the NanR-mediated gene regulation of sialic acid catabolism in *S. pneumoniae* (Figure 4.18). Here, three key states can be defined, including 1) free protein; 2) DNA-bound; and 3) effector-bound. The experimental evidence to support each state is summarised.

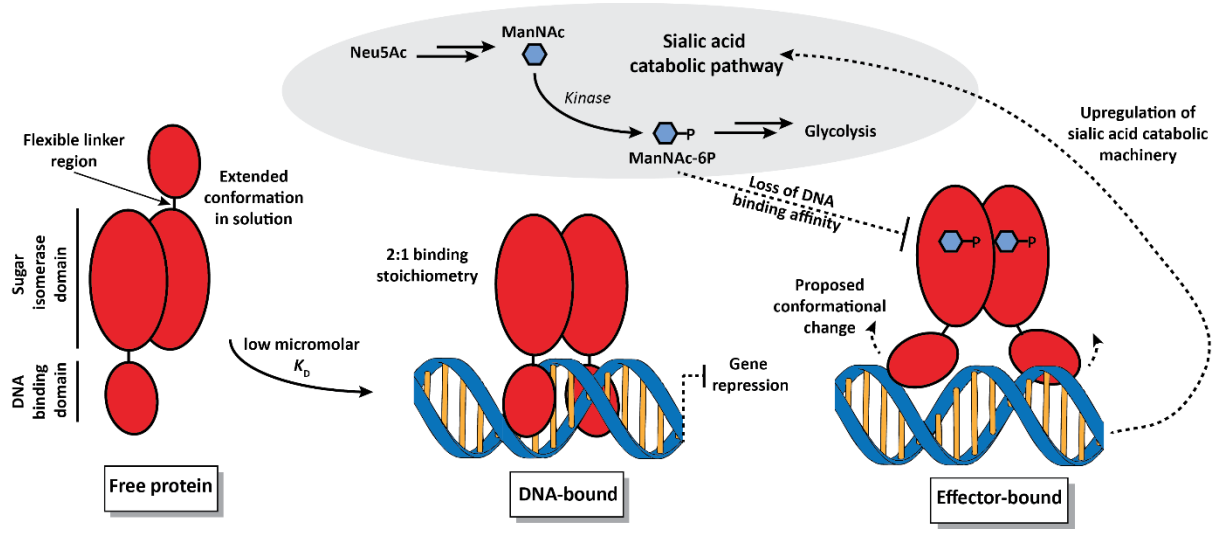


Figure 4.18 | Model mechanism of regulation for *S. pneumoniae* NanR. Briefly, the dimeric assembly of *S. pneumoniae* NanR (red) adopts an extended conformation in solution, when in the absence of DNA. In the presence of its DNA recognition sequence, NanR can coordinate DNA with low micromolar affinity forming a dimeric NanR-DNA hetero-complex. This DNA coordination is between the DNA-binding domains of NanR, as opposed to being oriented parallel to NanR. In this state, the expression of the sialic acid catabolic machinery is repressed. The binding of the sialic acid pathway catabolite *N*-acetylmannosamine-6-phosphate (ManNAc-6P) promotes a loss of DNA binding affinity. By consequence, a proposed conformational change leads to the upregulation of sialic acid catabolic machinery to promote glycolytic activity through the degradation of Neu5Ac.

Free protein. In the absence of a crystal structure, sedimentation velocity data demonstrated that *S. pneumoniae* NanR adopts a dimeric assembly in solution. Further SAXS experiments and evaluation of the solution structure with the theoretical scatter of *V. vulnificus* NanR supported that *S. pneumoniae* NanR adopts an extended conformation in solution. In this conformation, the N-terminal DNA-binding domains are oriented in opposing directions, mediated by a flexible linker, while the C-terminal sugar isomerase domain serves as the oligomeric interface.

DNA-bound. Sedimentation velocity studies demonstrated that *S. pneumoniae* NanR in the presence of DNA, coordinates DNA as a dimer with a 2:1 binding stoichiometry upon increasing protein concentration. Further analysis revealed a low micromolar DNA-binding affinity for its DNA recognition sequence. Structural studies using SAXS verified this stoichiometry and enabled a multiphase *ab initio* model to be reconstructed from the complex scattering data, along with the scattering data of the interaction partners (protein and DNA). Although low resolution, this model provided qualitative evidence to inform how NanR and DNA interact and suggested a conformational rearrangement of the N-terminal domains had occurred. This rearrangement is likely facilitated by the flexible linker region between the N- and C-terminal domains. A further comparison of this multiphase model to the

proposed 'active form' of *V. vulnificus* NanR highlighted an alternative shape in solution, supporting that the dimeric model of *V. vulnificus* NanR is incorrect.

Effector-bound. Differential scanning fluorimetry experiments verified the identity of the effector molecule, *N*-acetylmannosamine-6-phosphate following an increase in thermal stability. EMSA experiments demonstrated NanR can actively bind DNA, although in the presence of *N*-acetylmannosamine-6-phosphate this activity was significantly attenuated, further supporting its role as the effector molecule for *S. pneumoniae* NanR. Unfortunately, while diffraction data was collected for *S. pneumoniae* NanR in the presence of *N*-acetylmannosamine-6-phosphate to a resolution of 3.29-3.72 Å, this was not suitable for structure determination. These crystals should be optimised in the future in order to identify the residues that coordinate the effector molecule and subsequently map any conformational changes that have occurred as a result of effector binding.

Taken together, these data enhance our understanding for the RpiR-like sialoregulators and provide a platform for future structural characterisation at the atomic level to investigate effector binding and deduce specific protein-DNA contacts following crystallisation trials in the presence of DNA.

4.5 Chapter References

1. Jaeger, T., & Mayer, C. (2008). The transcriptional factors MurR and catabolite activator protein regulate *N*-acetylmuramic acid catabolism in *Escherichia coli*. *Journal of Bacteriology, 190*, 6598-6608.
2. Daddaoua, A., Krell, T., & Ramos, J. L. (2009). Regulation of glucose metabolism in Pseudomonas: the phosphorylative branch and entner-doudoroff enzymes are regulated by a repressor containing a sugar isomerase domain. *Journal of Biological Chemistry, 284*, 21360-21368.
3. Sorensen, K. I., & Hove-Jensen, B. (1996). Ribose catabolism of *Escherichia coli*: characterization of the *rpiB* gene encoding ribose phosphate isomerase B and of the *rpiR* gene, which is involved in regulation of rpiB expression. *Journal of Bacteriology, 178*, 1003-1011.
4. Afzal, M., Shafeeq, S., Ahmed, H., & Kuipers, O. P. (2015). Sialic acid-mediated gene expression in *Streptococcus pneumoniae* and role of NanR as a transcriptional activator of the *nan* gene cluster. *Applied and Environmental Microbiology, 81*, 3121-3131.
5. Johnston, J. W., Zaleski, A., Allen, S., Mootz, J. M., Armbruster, D., Gibson, B. W., Apicella, M. A., & Munson, R. S., Jr. (2007). Regulation of sialic acid transport and catabolism in *Haemophilus influenzae*. *Molecular Microbiology, 66*, 26-39.
6. Olson, M. E., King, J. M., Yahr, T. L., & Horswill, A. R. (2013). Sialic acid catabolism in *Staphylococcus aureus*. *Journal of Bacteriology, 195*, 1779-1788.
7. Zhu, Y., Nandakumar, R., Sadykov, M. R., Madayiputhiya, N., Luong, T. T., Gaupp, R., Lee, C. Y., & Somerville, G. A. (2011). RpiR homologues may link Staphylococcus aureus RNAIII synthesis and pentose phosphate pathway regulation. *Journal of Bacteriology, 193*, 6187-6196.
8. Therit, B., Cheung, J. K., Rood, J. I., & Melville, S. B. (2015). NanR, a Transcriptional Regulator That Binds to the Promoters of Genes Involved in Sialic Acid Metabolism in the Anaerobic Pathogen *Clostridium perfringens*. *Public Library of Science One, 10*, e0133217.
9. Bateman, A. (1999). The SIS domain: a phosphosugar-binding domain. *Trends in Biochemical Sciences, 24*, 94-95.
10. Brennan, R. G., & Matthews, B. W. (1989). The helix-turn-helix DNA-binding motif. *Journal of Biological Chemistry, 264*, 1903-1906.
11. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., & Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiology Reviews, 29*, 231-262.

12. Gourlay, L. J., Sommaruga, S., Nardini, M., Sperandeo, P., Deho, G., Polissi, A., & Bolognesi, M. (2010). Probing the active site of the sugar isomerase domain from *E. coli* arabinose-5-phosphate isomerase via X-ray crystallography. *Protein Science, 19*, 2430-2439.
13. Hwang, J., Kim, B. S., Jang, S. Y., Lim, J. G., You, D. J., Jung, H. S., Oh, T. K., Lee, J. O., Choi, S. H., & Kim, M. H. (2013). Structural insights into the regulation of sialic acid catabolism by the *Vibrio vulnificus* transcriptional repressor NanR. *Proceedings of the National Academy of Sciences of the United States of America, 110*, E2829-2837.
14. Cases, I., de Lorenzo, V., & Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends in Microbiology, 11*, 248-253.
15. Campilongo, R., Fung, R. K. Y., Little, R. H., Grenga, L., Trampari, E., Pepe, S., Chandra, G., Stevenson, C. E. M., Roncarati, D., & Malone, J. G. (2017). One ligand, two regulators and three binding sites: How KDPG controls primary carbon metabolism in Pseudomonas. *PLoS Genetics, 13*, e1006839-e1006839.
16. Vimr, E. R. (2013). Unified theory of bacterial sialometabolism: how and why bacteria metabolise host sialic acids. *International Scholarly Research Notices Microbiology, 2013*, 816713.
17. Almagro-Moreno, S., & Boyd, E. F. (2009). Sialic acid catabolism confers a competitive advantage to pathogenic *Vibrio cholerae* in the mouse intestine. *Infection and Immunity, 77*, 3807-3816.
18. Gualdi, L., Hayre, J. K., Gerlini, A., Bidossi, A., Colomba, L., Trappetti, C., Pozzi, G., Docquier, J. D., Andrew, P., Ricci, S., & Oggioni, M. R. (2012). Regulation of neuraminidase expression in *Streptococcus pneumoniae*. *BioMed Central Microbiology, 12*, 200.
19. Kim, B. S., Hwang, J., Kim, M. H., & Choi, S. H. (2011). Cooperative regulation of the *Vibrio vulnificus* nan gene cluster by NanR protein, cAMP receptor protein, and *N*-acetylmannosamine 6-phosphate. *Journal of Biological Chemistry, 286*, 40889-40899.
20. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
21. Parker, D., Soong, G., Planet, P., Brower, J., Ratner, A. J., & Prince, A. (2009). The NanA neuraminidase of *Streptococcus pneumoniae* is involved in biofilm formation. *Infection and Immunity, 77*, 3722-3730.
22. Kadioglu, A., Weiser, J. N., Paton, J. C., & Andrew, P. W. (2008). The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nature Reviews Microbiology, 6*, 288-301.
23. Almagro-Moreno, S., & Boyd, E. F. (2009). Insights into the evolution of sialic acid catabolism among bacteria. *BioMed Central Evolutionary Biology, 9*, 118.
24. Yesilkaya, H., Manco, S., Kadioglu, A., Terra, V. S., & Andrew, P. W. (2008). The ability to utilize mucin affects the regulation of virulence gene expression in *Streptococcus pneumoniae*. *FEMS Microbiology Letters, 278*, 231-235.
25. Trappetti, C., Kadioglu, A., Carter, M., Hayre, J., Iannelli, F., Pozzi, G., Andrew, P. W., & Oggioni, M. R. (2009). Sialic acid: a preventable signal for pneumococcal biofilm formation, colonization, and invasion of the host. *Journal of Infectious Disease, 199*, 1497-1505.
26. Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., Kazanov, M. D., Riehl, W., Arkin, A. P., Dubchak, I., & Rodionov, D. A. (2013). RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics, 14*, 745.
27. Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research., 14*, 1188-1190.
28. Moxon, E. R., & Murphy, P. A. (1978). Haemophilus influenzae bacteremia and meningitis resulting from survival of a single organism. *Proceedings of the National Academy of Sciences of the United States of America, 75*, 1534-1536.
29. Porterfield, J. Z., & Zlotnick, A. (2010). A simple and general method for determining the protein and nucleic acid content of viruses by UV absorbance. *Virology, 407*, 281-288.
30. Xiong, S., Zhang, L., & He, Q. Y. (2008). Fractionation of proteins by heparin chromatography. *Methods in Molecular Biology, 424*, 213-221.
31. Johnston, J. W., Shamsulddin, H., Miller, A.-F., & Apicella, M. A. (2010). Sialic acid transport and catabolism are cooperatively regulated by SiaR and CRP in nontypeable *Haemophilus influenzae*. *BMC Microbiology, 10*, 240-240.
32. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soeding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology, 7*.
33. Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., & Pabo, C. O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature, 298*, 447-451.

34. Schuck, P. (2013). Analytical Ultracentrifugation as a Tool for Studying Protein Interactions. *Biophysical Reviews, 5*, 159-171.
35. Cole, J. L., Lary, J. W., T, P. M., & Laue, T. M. (2008). Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. *Methods in Cell Biology, 84*, 143-179.
36. Schuck, P. (2000). Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophysical Journal, 78*, 1606-1619.
37. Jain, D. (2015). Allosteric control of transcription in GntR family of transcription regulators: A structural overview. *IUBMB Life, 67*, 556-563.
38. Balleza, E., López-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-Chávez, I., Balderas-Martínez, Y. I., Encarnación, S., & Collado-Vides, J. (2009). Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiology Reviews, 33*, 133-151.
39. Lebowitz, J., Lewis, M. S., & Schuck, P. (2002). Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein science : a publication of the Protein Society, 11*, 2067-2079.
40. Gorbet, G. E., Pearson, J. Z., Demeler, A. K., Colfen, H., & Demeler, B. (2015). Next-Generation AUC: Analysis of Multiwavelength Analytical Ultracentrifugation Data. In J. L. Cole (Ed.), *Analytical Ultracentrifugation* (Vol. 562, pp. 27-47).
41. Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols, 2*, 1849-1861.
42. Brookes, E., Cao, W. M., & Demeler, B. (2010). A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *European Biophysics Journal with Biophysics Letters, 39*, 405-414.
43. Demeler, B. (2005). UltraScan - A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments. In D. J. Scott, S. E. Harding, & A. J. Rowe (Eds.), *Analytical Ultracentrifugation: Techniques and Methods* (pp. 210-230): The Royal Society of Chemistry.
44. Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S., & Tucker, P. A. (2005). Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. *Acta Crystallographica Section D, 61*, 449-457.
45. Keegan, R. M., & Winn, M. D. (2007). Automated search-model discovery and preparation for structure solution by molecular replacement. *Acta Crystallographica D, Biological Crystallography, 63*, 447-457.
46. Walden, H. (2010). Selenium incorporation using recombinant techniques. *Acta Crystallographica Section D, Biological Crystallography, 66*, 352-357.
47. Mertens, H. D., & Svergun, D. I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology, 172*, 128-141.
48. Korasick, D. A., & Tanner, J. J. (2018). Determination of protein oligomeric structure from small angle X-ray scattering. *Protein science : a publication of the Protein Society, 27*, 814-824.
49. Pelikan, M., Hura, G. L., & Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. *General Physiology and Biophysics, 28*, 174-189.
50. Jacques, D. A., & Trewhella, J. (2010). Small-angle scattering for structural biology--expanding the frontier while avoiding the pitfalls. *Protein Science, 19*, 642-657.
51. Fischer, H., Neto, M. D., Napolitano, H. B., Polikarpov, I., & Craievich, A. F. (2010). Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *Journal of Applied Crystallography, 43*, 101-109.
52. Putnam, C. D., Hammel, M., Hura, G. L., & Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics, 40*, 191-285.
53. Brookes, E., Demeler, B., Rosano, C., & Rocco, M. (2010). The implementation of SOMO (SOlution MOdeller) in the UltraScan analytical ultracentrifugation data analysis suite: enhanced capabilities allow the reliable hydrodynamic modeling of virtually any kind of biomacromolecule. *European Biophysics Journal with Biophysics Letters, 39*, 423-435.
54. Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biochemical Journal, 76*, 2879-2886.
55. Petoukhov, M. V., & Svergun, D. I. (2015). Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Crystallographica Section D, Biological Crystallography, 71*, 1051-1058.
56. Svergun, D., Barberato, C., & Koch, M. H. (1995). CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography, 28*, 768-773.
57. Boldon, L., Laliberte, F., & Liu, L. (2015). *Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application* (Vol. 6).

Chapter Five

Defining the DNA-binding mechanism of the GntR-type sialoregulator from *Escherichia coli* – Part One

The work presented in this chapter was a collaborative project with both Professor Borries Demeler at the University of Lethbridge, Canada and Dr Hariprasad Venugopal at the Clive and Vera Ramaciotti Centre for Cryo-Electron Microscopy, Monash University, Australia. Over a total of five weeks I visited the University of Lethbridge to conduct novel multi-wavelength sedimentation velocity experiments to study the interactions between *Escherichia coli* NanR and its DNA recognition sequence in the solution state. With guidance from Professor Borries Demeler, I developed an understanding towards the experimental design, data analysis, and result interpretation for this novel biophysical technique. This research was funded by a grant from the Maurice Wilkins Centre as this advanced analytical ultracentrifugation instrumentation is not available in New Zealand. Secondly, I visited the Clive and Vera Ramaciotti Centre for Cryo-Electron Microscopy several times, where I undertook the work presented in this chapter with guidance from Dr Hariprasad Venugopal. Specifically, I assisted in the sample preparation and initial screening of the cryo-electron microscopy data, while Hariprasad processed the data and successfully reconstructed the structure of the NanR-DNA hetero-complex. This structure is reported in the manuscript within this Chapter, which I analysed, wrote and produced the figures.

5.1 Introduction

5.1.1 GntR family of transcriptional regulators

First discovered in *Bacillus subtilis,* GntR is a protein that acts as a transcriptional repressor of the gluconate (*gnt*) operon (44) and is the defining member of the GntR superfamily. Today, the GntR superfamily is one of the most widespread groups of bacterial transcriptional regulators, which regulate a variety of key biological processes within bacteria (45).

Members of this regulatory superfamily are structurally comprised of an N-terminal DNA-binding domain, which contains a winged helix-turn-helix motif, and a C-terminal effector-binding domain (46; 47). As described in Chapter One, the helix-turn-helix motif of GntR members is characterised by an additional 'wing' that interacts with the minor groove of DNA to provide additional specificity. This 'wing' is formed by a small loop with two short β-strands following the canonical tri-helical structure (48). Together, the architecture of this DNA-binding motif is highly conserved among GntR members. Although surprisingly, the sequence similarity of the DNA-binding motif between GntR members is low (~25%). In contrast, the C-terminal domain exhibits a variety of different structural folds (45), although they share a common function in effector binding and forming the oligomerisation interface. These structural folds can be divided into seven subfamilies based on shared structural features: AraR; DevA; FadR; HutC; MocR; PlmA; and YtrA, (45; 49-53). Together, these seven subfamilies regulate an array of biological processes in bacteria, including various metabolic or biosynthetic pathways, the activity of the ABC transport systems, and as environmental sensors that respond to sources of nutrition (45-47). All subfamilies, except DevA and PlmA have representative structures in the PDB, presented in Figure 5.1.

Between these subfamilies, the FadR subfamily is the biggest, comprising about 40% of all GntR family members (47). The first member to be identified in this subfamily was the repressor FadR, which regulates the genes responsible for fatty acid metabolism (54). In addition, this repressor is also considered a model protein of the GntR superfamily with numerous crystal structures determined from *E. coli* and *V. cholerae*. Notably, these include the apo transcriptional regulator (55) and hetero-complex structures of FadR with DNA of the *fad* operon (56) and with various fatty acid effector molecules (57; 58).

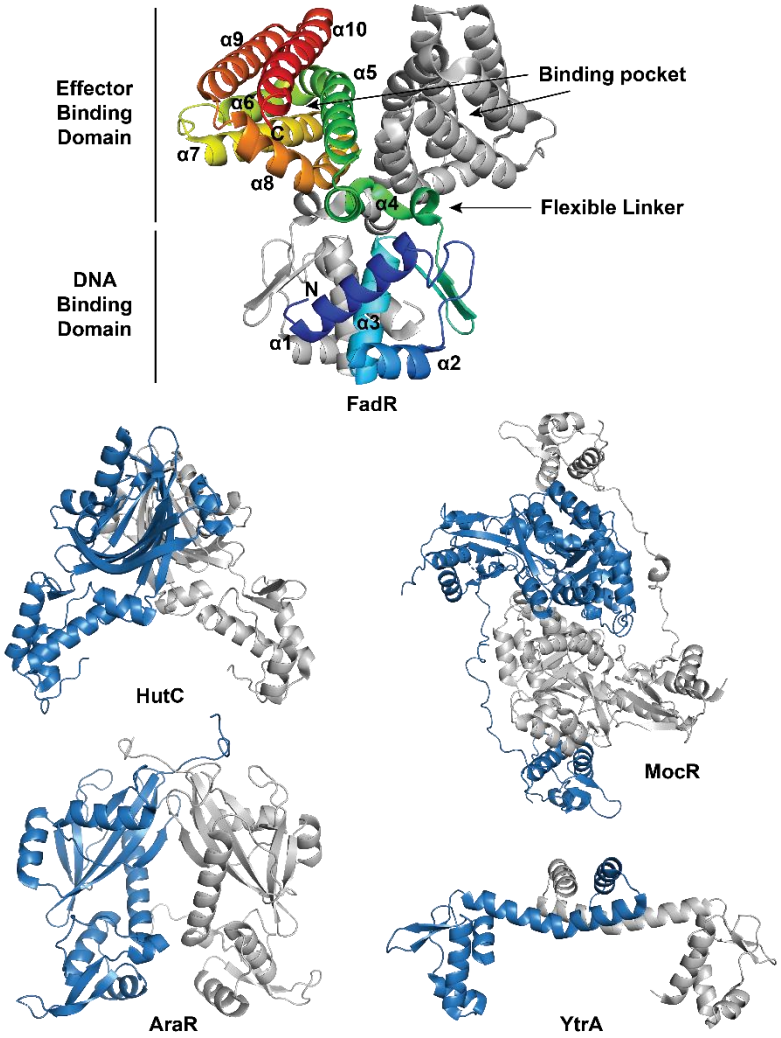


Figure 5.1 | The structural architecture of the GntR subfamilies. The domain architecture of the largest subfamily, FadR is presented. Shown is the homodimer (PDB ID - 1HW1) with one monomer colour-coded and labelled, while the other is coloured grey. Representative homo-dimeric structures of the remaining subfamilies are presented: HutC subfamily (PDB ID - 4U0V); MocR subfamily (PDB ID - 4N0B); AraR subfamily (PDB ID - 5DEQ); and the YtrA subfamily (PDB ID - 2EK5). For each, one monomer is coloured in blue, while the other is coloured in grey. The topology of the N-terminal DNA-binding domain is very similar, while the C-terminal domain architecture is very different between subfamilies.

Structurally, FadR members are characterised by an all α-helical C-terminal domain, which forms a tight antiparallel bundle. Depending on the number of α-helices in this helical array, the subfamily can be further classified into two subgroups: VanR, which is comprised of six α-helices; and FadR, which is comprised of seven α-helices. This additional helix serves as a flexible linker connecting the N-terminal DNA-binding domain to the C-terminal effector-binding domain, as opposed to a short loop in the VanR subgroup (45). In some cases, this additional helix can facilitate a domain swap through the C-terminal domain (59).

Interestingly, members of this subfamily, such as TM0439 from *Thermotoga maritima* (51) are metal-dependent, coordinating various metal ions such as zinc, through three conserved histidine residues buried deep within the C-terminal domain (51; 59). Although, the role these metal ions play is currently unclear. Do metal ions play a structural role and facilitate the binding of the effector molecule or do they mediate oligomerisation? Alternatively, are these metal-sensing transcriptional regulators? Further structural insight is required to fully understand the physiological role these metal ions provide select repressors, whom some authors believe should form a new subgroup within the FadR subfamily (51).

5.1.2 Interaction of GntR family members with DNA

Identified through protein-DNA co-crystal structures, GntR members are known to interact with DNA as dimers, where the recognition helix (α3) of the winged helix-turn-helix motif binds within the major groove, perpendicular to the double-stranded DNA axis. Here, residues at the end of the recognition helix can directly interact with specific bases through electrostatic interaction, or by forming a hydrogen-bonding network (56; 60; 61). For example, the identification of a conserved arginine residue present on the recognition helix of the winged helix-turn-helix motif interacts directly with a guanine base through hydrogen bonding (46). Conversely, the 'wing' of the DNA-binding motif interacts with the minor groove of DNA. Interestingly, a conserved glycine residue present at the tip of the 'wing' has been identified as a key interaction within the minor groove. Mutation of this key glycine residue to introduce steric bulk or charge significantly affects the ability of the repressor to bind DNA (62). Consequently, this glycine residue is a trademark of this protein family.

The point of interaction for each monomer can be described as a half-site, which typically represents a perfect- or pseudo-palindromic sequence that is rich in adenine and thymine bases (47). The centre of these palindromic half-sites is highly conserved, while the periphery is divergent and thus contributes to specificity of the repressor (63). Although, in contrast some GntR members recognise inverted or direct repeats that offer no symmetry (45; 60). This is the case for the GntR-type NanR from *Escherichia coli*. Unfortunately, protein-DNA hetero-complex structures have only been solved for GntR members that bind palindromic sequences (56; 60; 64). The interaction with a direct repeat has yet to be characterised at the molecular level. However, the structural data available still provides invaluable insight into the molecular recognition at the DNA interface. A representative model of the DNA bound conformation is presented in Figure 5.2.

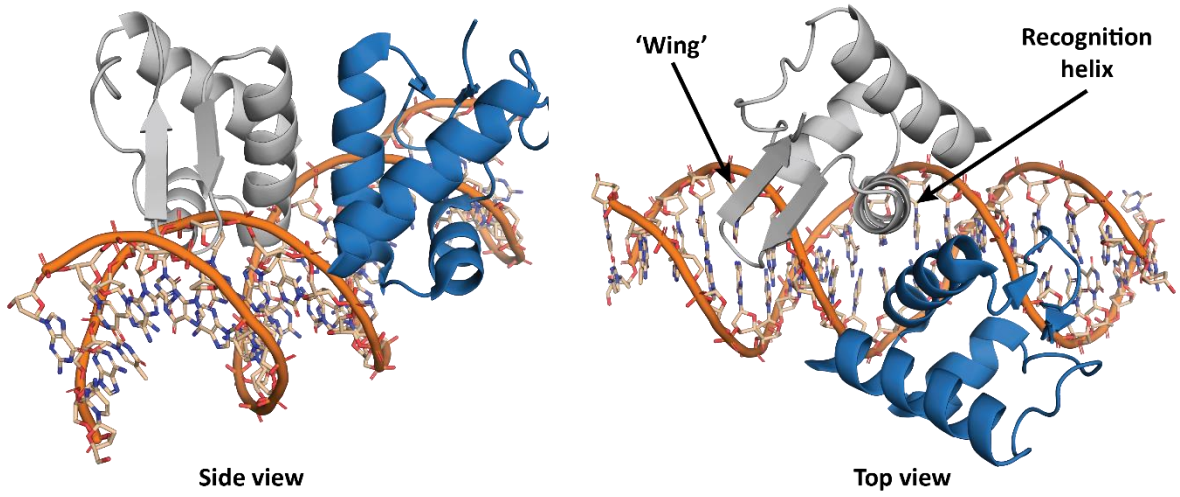


Figure 5.2 | The interaction of the N-terminal DNA-binding domain with DNA. The structure of the *E. coli* FADR N-terminal DNA-binding domain is presented (PDB 1HW2) as a representative model of the GntR superfamily. All other solved complex structures bind DNA in a similar manner. One monomer is coloured blue and the other is coloured grey. Highlighted is the 'wing', which interacts with the minor groove of DNA and the recognition helix, which interacts with the major groove of DNA.

Additionally, the distance between the half-sites is crucial for presenting the binding site in the correct orientation for the repressor and therefore contributes to the DNA binding mechanism. This process can be intimately associated with the flexibility of the linker region between the N- and C-terminal domains (47) (Figure 5.1, top panel). Here, small distances facilitate tight and compact interaction with DNA, while much longer distances can facilitate DNA looping (45). Additional features that may influence DNA binding are yet to be discovered.

5.1.3 Allosteric regulation mechanism of GntR family members

Within the GntR superfamily, the allosteric mechanism of regulation has been characterised in *E. coli* FadR (55), *Vibrio cholerae* FadR (58), and *Bacillus subtilis* HutC (61; 65). Here, upon effector-binding, a large displacement of the N-terminal DNA-binding domains leads to an attenuation of DNA binding activity, a process mediated by the flexible linker that connects the N- and C-terminal domains. Interestingly, this displacement of the DNA-binding domains has been likened to a 'jumping-jack' motion in *B. subtilis* HutC in which the entire domain moves an astonishing 44 Å (65). This movement is similar in both FadR repressors, although less pronounced (58; 61).

To fully understand the allosteric behaviour of these transcriptional regulators, further structural data is required, along with a greater understanding of molecular dynamics that are involved. As a result, this may uncover the potential for these regulators to serve as viable drug targets, such as the tetracycline

repressor in *E. coli* (66). Here, effector-screening has identified novel compounds, aside from their natural effector that can enhance the DNA-binding affinity and some that can reduce the affinity (67). This feature may provide insight to specifically engineer repressors for therapeutic application (68; 69).

5.1.4 Regulation of sialic acid catabolism in *Escherichia coli*

Given the importance of sialic acid to the survival and persistence of bacteria (as discussed in Chapter One), it comes as no surprise that its catabolic pathway is regulated to ensure an efficient metabolic return of this alternative source of nutrition. In *E. coli*, gene regulation of the core catabolic operon, *nanATEK-yhcH* operon is controlled by the GntR-type transcriptional repressor NanR (70).

To achieve regulation, NanR binds a conserved operator site that contains three exact repeats of the DNA sequence GGTATA (non-palindromic). This hexanucleotide sequence, defined as the *N*-acetylneuraminic acid (Nan) box is located within the intergenic region between the *nanR* gene and *nanATEK-yhcH* operon (71). In addition to regulating the core catabolic machinery for sialic acid, *E. coli* NanR regulates the gene expression of the *nanCMS* operon, encoding the proteins responsible for outer-membrane transport and periplasmic processing of sialic acid (see Chapter One, Section 1.5.1), and the *yjhBC operon*, which was investigated in Chapter Two and Three. Together, these operons also contain three exact or near-exact repeats of the DNA sequence GGTATA (72).

Collectively, these ten genes that are regulated by NanR form the sialoregulon. Interestingly, this coordinated regulation is only present in *E. coli*, *Shigella dysenteriae* and select strains of Shiga-toxin-producing *E. coli* (73). This unique feature has made *E. coli* a model organism for understanding sialic acid catabolism, among bacteria who utilise this alternative source of nutrition (74). An overview of the *E. coli* sialoregulon along with the DNA-binding sequence for each regulated operon is presented below in Figure 5.3. While none of these operons are clustered together in the genome, the *yjhBC* (Base pair 4503624-4506405) and *nanCMS* (Base pair 4540216-4536614) operons are localised much closer to each other than the core *nanATEK-yhcH* operon (Base pair 3376991-3360914).

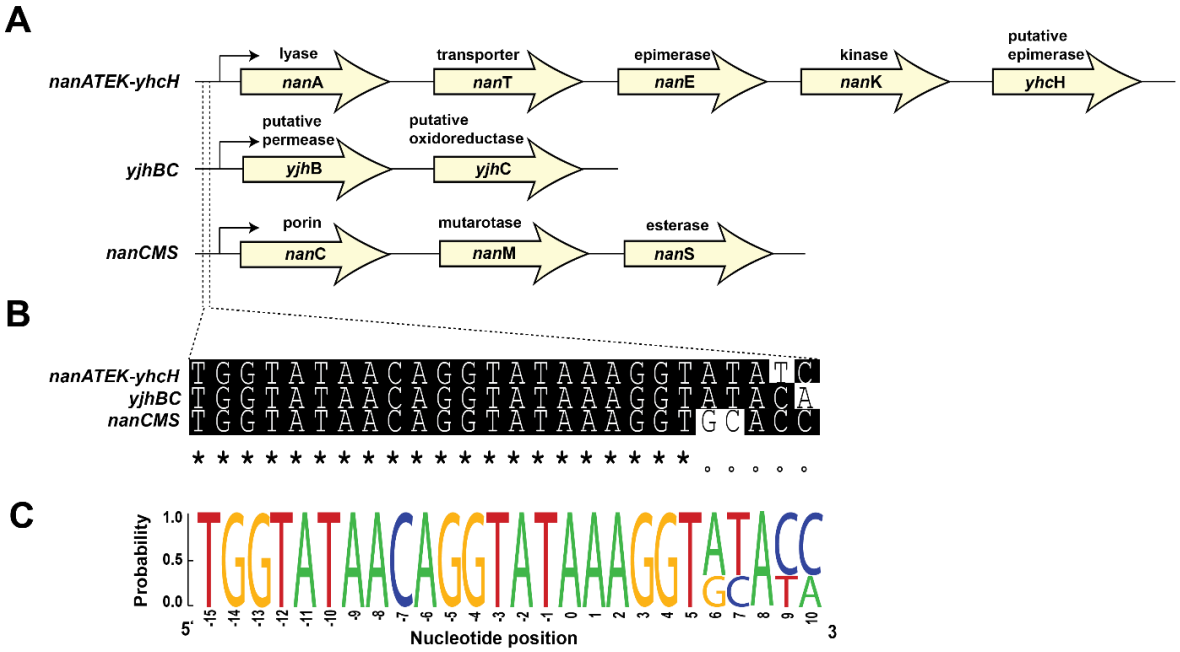


Figure 5.3 | The *E. coli* Sialoregulon. A) Together the *nanATEK-yhcH*, *yjhBC* and *nanCMS* operons define the sialoregulon, regulated by the transcriptional regulator NanR. **B)** The NanR DNA-binding sequence for each operon is presented. This is located at the operator site for each respective operon. **C)** A sequence logo of each respective operator site, where the overall height indicates the sequence conservation at that position. This highlights that NanR binds a conserved operator site that contains three exact or near-exact repeats of the DNA sequence GGTATA. Figure generated using WebLogo (75).

5.1.5 Previous studies on *E. coli* NanR.

The current understanding of *E. coli* NanR in the literature has largely been influenced by the studies of Kalivoda and colleagues (71; 72). Notably, through DNA footprint analysis this has included the identification that the sequence GGTATA is the motif recognised by NanR, and when subjected to an electrophoretic mobility shift assay (EMSA), three concentration-dependent complexes are formed. It is believed that the discrete complexes represent the binding of sequential NanR promoters to the operator site on DNA and not conformational isomers. However, evidence from a Ferguson analysis, a gel-based technique to estimate the size of macromolecules was largely inconsistent (72), therefore the stoichiometry of this interaction is poorly understood. This is likely reflective on the inherent limitations with EMSA, such as the exposure to a non-equilibrium environment and unnatural stabilisation in the gel—a phenomenon that is poorly understood, but hypothesised to be the result of macromolecular crowding (76). Moreover, the use of extrinsic labels to detect complexes by EMSA may affect DNA-binding activity (76; 77).

Neu5Ac is implicated as an effector molecule *in vivo* following a 10-fold induction of *nan* genes using a *lacZ* translational fusion assay. This however, has not been demonstrated *in vitro* using EMSA, but the presence of Neu5Ac is reported to dissociate NanR promoters into monomers (72). This feature is believed to be intimately linked with the regulation mechanism (72). In addition, DNA methylation is believed to contribute to the full induction of the sialoregulon, as located immediately upstream of the first direct GGTATA repeat is a DAM methylase GATC consensus motif (74). Here, it is hypothesised that once NanR dissociates from the operator site, DAM methylation may reduce the repressors ability to rebind as the effector concentration decreases in the cell, leading to complete catabolism of sialic acid (78; 79).

Despite this largely genetic characterisation for *E. coli* NanR, the biggest gap in the field is the paucity of structural or biophysical data to dissect the allosteric mechanism, identify how the effector molecule coordinates the C-terminal, and elucidate how the GntR-type NanR binds DNA.

5.2 Chapter overview

The main aim of this chapter is to develop a mechanistic understanding of how *E. coli* NanR regulates gene expression. Reported as a manuscript, this body of work quantitatively defines the stoichiometry of the *E. coli* NanR DNA complexes, label-free and in solution using multi-wavelength sedimentation velocity experiments. The binding kinetics report a nanomolar DNA binding constant and reveal novel cooperative behaviour between the promoters that bind the DNA. The first crystal structure of a GntR-type sialoregulator is reported, solved in complex with Neu5Ac and zinc, identifying a metal-ligand binding pocket. Excitingly, using single particle cryo-electron microscopy I report a structure of the NanR-DNA hetero-complex, which provides novel insight into how NanR binds DNA and additionally provide structural evidence to support the stoichiometry that was observed during multi-wavelength sedimentation velocity experiments.

Combined, this analysis provides the first molecular insight into GntR-type sialoregulators and enhances our understanding of how the GntR family of transcriptional regulators control gene expression.

5.3 Manuscript

Title:

The cooperative assembly and disassembly of the GntR-type sialoregulator NanR from *Escherichia coli*.

Christopher R. Horne [a, #], Hariprasad Venugopal [b, #], Santosh Panjikar [c, d], Amy Henrickson [e], Rachel A. North [a], Michael D. W. Griffin [f], Georg Ramm [b], Borries Demeler [e], Renwick C.J. Dobson [a, f, *]

[a] Biomolecular Interaction Centre and School of Biological Sciences, University of Canterbury, PO Box 4800, Christchurch 8140, New Zealand.

[b] Clive and Vera Ramaciotti Centre for Cryo-Electron Microscopy, Monash University, Clayton, Victoria 3800, Australia.

[c] Australian Synchrotron, Clayton, Victoria 3168, Australia.

[d] Department of Molecular Biology and Biochemistry, Monash University, Melbourne, 3800, VIC, Australia.

[e] Department of Chemistry and Biochemistry, University of Lethbridge, Lethbridge, Alberta T1K 3M4, Canada.

[f] Bio21 Molecular Science and Biotechnology Institute, Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria 3010, Australia.

[#] *First authors*

** Corresponding author*

Keywords: NanR, transcriptional regulator, GntR superfamily, sialic acid, Neu5Ac, cooperativity, X-ray crystallography, Analytical ultracentrifugation, Cryo-electron microscopy

Abstract:

Bacteria carefully regulate metabolic processes to efficiently colonize and persist within the human host. One mechanism bacteria employ is gene regulation. In *Escherichia coli,* the catabolism of sialic acid is controlled by the GntR-type transcriptional regulator NanR, but the mechanism of gene regulation is unknown. We first demonstrate that NanR binds as a dimer to a total of three GGTATA direct repeats that make up the DNA recognition site, forming a hexameric assembly. Interestingly, we found this binding is cooperative and demonstrate it is mediated by a unique N-terminal extension, likely through protein-protein interactions. To understand how DNA-binding is attenuated by the effector, we solve the co-crystal structure of *E. coli* NanR in the presence of *N*-acetylneuraminic acid to 2.1 Å, supporting the hypothesis that this is the effector molecule, and identify a metal-binding motif that coordinates zinc. The structure is asymmetrical with *N*-acetylneuraminic acid and zinc only present in one monomer, informing the molecular details of the conformational change that occurs following binding to attenuate DNA-binding activity. Notably, we report the structure of the NanR-DNA hetero-complex to 3.9 Å using cryo-electron microscopy, and the first structural evidence of a multimeric assembly process within the GntR superfamily. Together, our results give the first molecular insight into the mechanism of the NanR-DNA interaction in *E. coli,* which enhances our understanding of sialic acid gene regulation in bacteria.

Significance Statement:

To gain a competitive advantage in the human host, pathogenic bacteria scavenge and degrade sialic acid as an alternative source of carbon, nitrogen and energy. The expression of the genes, responsible for the utilization of this amino sugar are controlled by the repressor protein, NanR. How NanR interacts with DNA to regulate this process is poorly understood at a molecular level. Here, we report a structural and functional characterization of NanR and define how the protein-DNA complex is formed. Based on these data, we propose a model for regulation, showing that NanR binds DNA cooperatively and describe how this interaction may be attenuated by the binding of sialic acid and the metal ion zinc.

Introduction

The human gastrointestinal tract is one of the most heavily populated niches occupied by bacteria (1). This environment is largely glucose-limited (2; 3) and gain a competitive advantage, bacteria must rapidly adapt to nutrient changes, a response that is facilitated by transcriptional regulators at the molecular level (4; 5). In *Escherichia coli*, the NanR repressor regulates the expression of genes responsible for sialic acid uptake and catabolism (6) (Fig. 1). Sialic acids are a family of negatively-charged, nine-carbon amino monosaccharide units located at the terminal end of glycoconjugate structures found in the mucosal epithelia of the mammalian respiratory and gastrointestinal tract (7; 8). Although there are more than 50 naturally occurring and structurally distinct sialic acid variants, the most ubiquitous form across all organisms is *N*-acetylneuraminic acid (or Neu5Ac) (Fig. 1*A*) (9; 10). In the human host, Neu5Ac serves as a source of carbon, nitrogen and energy for pathogenic and commensal bacteria that have evolved catabolic machinery to utilize this alternative source of nutrition (10; 11).

As a negative regulator, *E. coli* NanR binds with high affinity to a conserved DNA recognition site containing three exact (or near exact), GGTATA direct repeats (6; 12). This hexanucleotide motif, designated the *N*-acetylneuraminic acid (Nan) box is present in three distinct operons (Fig. 1*B-C*): 1) *nanATEK-yhcH*, the core catabolic pathway (6; 11); 2) *nanCMS*, responsible for outer-membrane transport and periplasmic processing (13-15); and 3) *yjhBC*, encoding proteins of unknown function, yet hypothesized to process less common variants of sialic acid (11). Collectively, the coordinated regulation of these 10 genes by NanR is referred to as the sialoregulon, and has only been identified in *E. coli*, *Shigella dysenteriae* and some Shiga-toxin-producing *E. coli* strains (11; 12). *E. coli* NanR belongs to the GntR superfamily of transcriptional regulators. These proteins are composed of an N-terminal DNA-binding domain, which contains a winged-helix-turn-helix (wHTH) motif, and a C-terminal effector-binding domain. While the architecture of the wHTH DNA-binding domain is highly conserved (16), the C-terminal effector-binding exhibits a wide variety of structural folds—these can be divided into seven subfamilies based on shared structural features (17). To attenuate repression, an effector molecule binds to GntR regulators, causing a conformational change that alters DNA binding affinity (18). This allosteric regulation is a trademark of bacterial transcriptional regulators allowing them to function as molecular switches to turn gene expression on or off as required (19). In *E. coli* NanR, Neu5Ac is thought to be the effector molecule and inducer of the sialocatabolic pathway *in vivo*, converting NanR oligomers to monomers in solution (19). However, the molecular mechanism that facilitates this allosteric regulation *in vitro* is poorly understood, with no direct biophysical or structural evidence confirming that Neu5Ac affects the protein-DNA interaction.

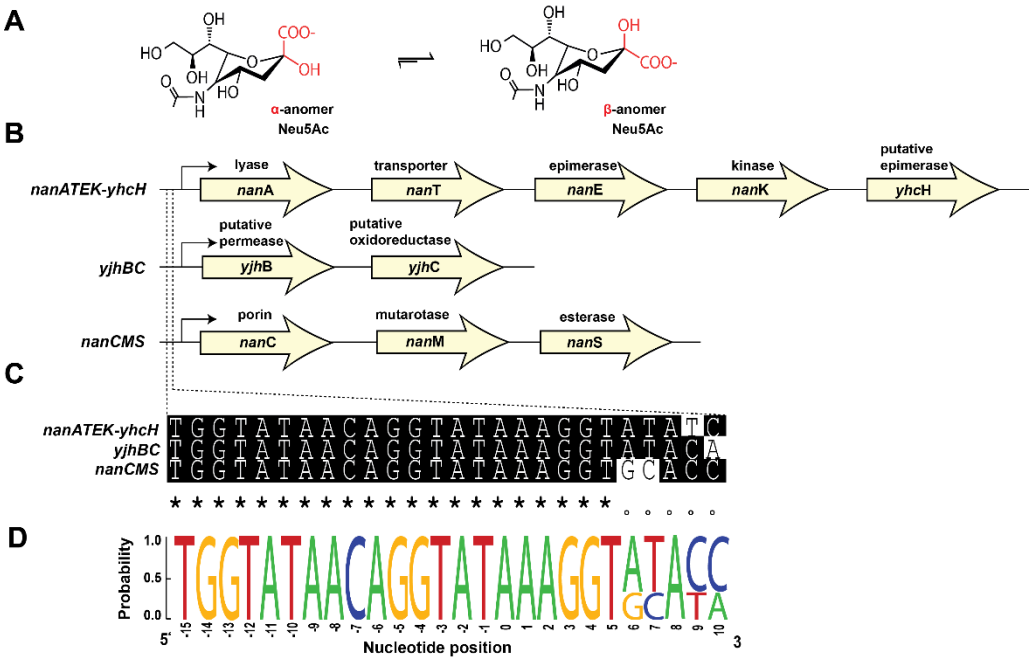


Fig. 1 | The transcriptional repressor NanR binds a conserved operator site in three distinct operons, collectively referred to as the sialoregulon, to coordinate the catabolism of sialic acid in *E. coli*. (***A***) The most common sialic acid, Neu5Ac is shown in chair conformation. Red indicates the stereochemical difference at the C2 position, generated the α- or thermodynamically favorable β-anomer. (***B***) *nanATEK-yjhcH*, *yjhBC* and *nanCMS* operons within the sialoregulon. (***C***) Sequence alignment of the operator sites present in each operon. Conserved nucleotides are marked with an asterisk. (***D***) Based on this alignment, this sequence logo, generated using WebLogo (20) highlights the conservation of the DNA bases within these operator sites.

Here, we demonstrated that complex formation is cooperative and mediated by protein-protein interactions. We then defined the stoichiometry of the protein-DNA hetero-complex, label-free and in solution, *via* multi-wavelength sedimentation velocity experiments. We also report a 2.1 Å co-crystal structure of *E. coli* NanR in the presence of Neu5Ac, which reveals a metal-ligand binding pocket. Using cryo-electron microscopy, we present a 3.9 Å structure of the NanR-DNA hetero-complex (70.5 kDa) and provide structural insight to support the formation of the higher-order hetero-complex. Taken together, our biophysical and structural data give the first molecular insight into the mechanism of the NanR-DNA interaction in *E. coli* and enhance our understanding of how GntR-type regulators control gene expression.

Results and Discussion

NanR binds DNA cooperatively with nanomolar affinity.

The presence of multiple binding sites suggests the mechanism for gene regulation involves cooperativity, as observed for the well-known *lac* (21) and *ara* (22) repressors through DNA-looping and

in other GntR family members such as CitO (23) and PhnF (24). To determine whether *E. coli* NanR exhibits cooperative behavior, we titrated NanR against fluorescently-labelled DNA oligonucleotides that mimic the full recognition site, by employing both an electrophoretic mobility shift assay (EMSA) (Fig. 2*A*) and analytical ultracentrifugation (Fig. 2*B*). As shown in Fig. 2*C*, NanR binds three repeats of the GGTATA sequence with nanomolar affinity, displaying cooperative behavior, as shown by the sigmoidal dependency and fit to the Hill model. EMSA experiments were conducted using a nanomolar titration range of NanR against a set concentration of fluorescently end-labelled DNA. As visualized using a fluorescent scanner, three concentration-dependent complexes were formed. Densitometry analysis of the observed bands then revealed the ratio of bound and unbound DNA. Fitting this data to the Hill equation gave a sigmoidal binding isotherm (Fig. 2*C*), with a Hill coefficient of 2.5 ± 0.3, which is characteristic of cooperativity. The apparent dissociation constant (K_D) of 36.3 ± 1.8 nM is consistent with previous reports (6; 12) and is within range of the reported K_D values of other regulators within the GntR superfamily (25). At high protein concentrations, we observed the development of non-specific binding, a feature also detected in the poly-AT oligonucleotide control, where specific binding was abrogated (*SI Appendix*, Fig. S1*A*). This non-specific binding is likely associated with large protein-DNA hetero-complexes, mediated by the localized concentration of NanR, as these species are very poorly resolved in the gel shift assay and do not substantially affect the signal of the DNA probe. For clarity we also assessed DNA binding by EMSA with oligonucleotides containing one or two GGTATA repeats (*SI Appendix*, Fig. S1*B*-*C*). In these experiments, complex formation was only observed by EMSA with two GGTATA repeats. This observed loss of DNA binding activity with only one GGTATA repeat further supports the hypothesis that the mechanism of regulation for NanR depends on cooperative binding, because multiple GGTATA repeats appear essential to both initiate complex formation and drive the formation of the higher-order assemblies.

We also sought to assess the DNA binding kinetics of NanR more quantitatively and without the limitations of a gel-based technique by employing analytical ultracentrifugation. Here, using fluorescence optics, sedimentation velocity experiments were conducted by using an identical NanR titration range against the fluorescently-labeled DNA and measurement at an emission wavelength of 495 nm. This enabled detection of free DNA and an evaluation of protein-DNA complex formation indicated by a positive shift in the sedimentation coefficient. When fitted to a c(*s*) model, three populations of concentration-dependent complexes (C1-3) were visualized in the c(*s*) distribution across the NanR titration range (Fig. 2*B*)—this mirrors what we observed by EMSA (Fig. 2*A*). To determine the binding kinetics, the ratio of bound and unbound DNA was determined by peak integration within the c(*s*) distribution. When plotted and fit to the Hill model, the binding isotherm also exhibited a sigmoidal shape (Fig. 2*C*), with a Hill coefficient of 1.9 ± 0.2 and an apparent K_D of 19.7 ± 1.3 nM, consistent with

cooperative binding. Collectively, these DNA-binding studies support the hypothesis that the coordinated regulation of the *E. coli* sialocatabolic pathway is driven by nanomolar affinity and mediated by positive cooperativity. Insight into what drives this cooperativity will be discussed later in this study.

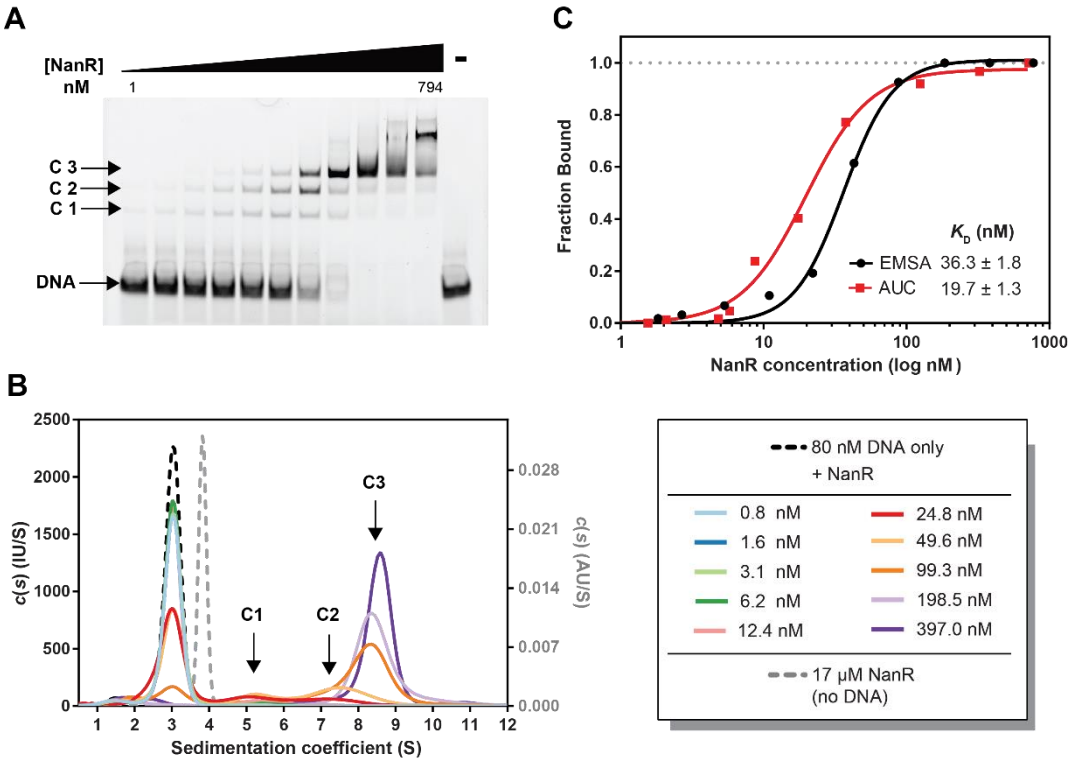


Fig. 2 | *E. coli* NanR binds DNA with nanomolar affinity and displays positive cooperativity. (***A***) Using fluorescently labelled DNA and titration of NanR, an electrophoretic mobility shift assay highlights that NanR binds DNA, resulting in three concentration-dependent complexes (C1–3). (***B***) Sedimentation velocity data was fitted to a continuous mass [c(*s*)] distribution. With increasing NanR concentration, the DNA signal at ~3 S slowly shifts, consistent with complex formation. The three populations of concentration-dependent complexes (C1–3), analogous to the EMSA experiment, are highlighted. The pure protein control, run separately, is shown for reference (grey dashed line). (**C**) The binding isotherm from the EMSA (black) and AUC (red) data was best fit to the Hill equation.

E. coli NanR remains dimeric in solution, regardless of Neu5Ac.

Validation of the oligomeric assembly is essential for developing a functional understanding of how transcriptional regulators operate. Most members of the GntR superfamily function as a dimer in solution to control gene expression, although tetrameric assemblies have been characterized for some members (26; 27). Moreover, it has been reported than Neu5Ac, implicated as the effector molecule of *E. coli* NanR *in vivo,* induces a dramatic dissociation in SDS-PAGE analysis, converting NanR oligomers to monomers (12). Together, these observations led us to evaluate the oligomeric state of *E. coli* NanR and

investigate Neu5Ac-induced dissociation in solution using single-wavelength sedimentation velocity and small angle X-ray scattering (SAXS) experiments. Both demonstrated that *E. coli* NanR is a dimer in solution, regardless of Neu5Ac.

Sedimentation velocity data were fitted to a continuous mass distribution [c(*M*)] model, resulting in a single, monodisperse peak for both samples. The c(*M*) distribution indicated an apparent molar mass for NanR with (55.6 kDa) and without (54.5 kDa) Neu5Ac that were consistent with the calculated dimeric mass of 59 kDa for NanR (*SI Appendix*, Fig. S2*A*). Further fitting to a continuous sedimentation coefficient distribution [c(*s*)] model was also consistent with a single oligomeric species in solution for both samples (*SI Appendix*, Table S1). SAXS data were also consistent with a dimeric NanR assembly in solution when analyzed using SAXS-MoW2 (28). The predicted molar mass values for NanR with (63.6 kDa) and without (70.6 kDa) Neu5Ac were consistent with the calculated dimeric mass, despite the slightly larger size of NanR in the absence of Neu5Ac. In addition, both samples presented an analogous scattering profile (*SI Appendix*, Fig. S1*B*), further supporting a single oligomeric species in solution. Taken together, these data propose a dimeric assembly of NanR in solution that is not perturbed by the presence of Neu5Ac, contradicting reports that Neu5Ac-induced dissociation is intimately linked with the mechanism of regulation. All data collection and data analysis statistics for these experiments are summarized in *SI Appendix*, Table S1.

Stoichiometry of the NanR-DNA complex.

Understanding the stoichiometry of protein-DNA hetero-complexes is crucial to elucidating the DNA-binding mechanism. In this system, DNase footprint analysis, conducted by Kalivoda et al. revealed that NanR accommodates the full recognition site by overlapping individual GGTATA repeats, whereby one NanR promoter provides coverage of one GGTATA repeat, as well as partial coverage of a neighboring GGTATA repeat, approximately a half turn away on the DNA helix (6). Furthermore, initial binding has been postulated to involve a homotrimer of NanR, with full coverage achieved by six monomers over a region of approximately 30 bp, or three turns of the DNA helix (6; 12). This is inconsistent with both our single-wavelength sedimentation velocity and small angle X-ray scattering (SAXS) experiments, which proposed a dimeric species of NanR in solution (*SI Appendix*, Fig. S2).

Thus, we investigated the nature of the NanR-DNA interaction, label-free and in solution by multi-wavelength sedimentation velocity (MWL-SV) experiments. This emerging analytical ultracentrifugation strategy couples a novel multi-wavelength detector with high-performance computing to characterize complex mixtures by deconvoluting the spectral signals of the interaction partners into separate sedimentation profiles. As an intrinsic extinction profile of each interaction partner is used to achieve decomposition, the separate sedimentation profiles are scaled to molar concentration. This is facilitated

by employing the non-negatively-constrained least-squares algorithm (29; 30). This feature allows the stoichiometry of the complex to simply be extracted by integrating the molar ratio of the co-migrating peaks (31). In addition, the separated datasets can then be further analyzed to extract hydrodynamic information as is traditionally achieved using single-wavelength experiments. Thus, this new approach provides both hydrodynamic and spectral characterization of an interacting system to define the stoichiometry of association, and grant access to properties such as the mass and frictional ratio of each species. Together, this enables interacting species to be characterised with high resolution.

Using the DNA oligonucleotide with three GGTATA repeats as a model of the full operator site, we observed three concentration-dependent complexes (Fig. 2). However, this suggests a complex system involving multiple interaction partners. While our overarching aim is to define the stoichiometry of protein-DNA complex formation in the full recognition site, we first characterized the assembly process with only two GGTATA repeats present. EMSA revealed the formation of only two complexes (*SI Appendix*, Fig. S1*C*), thus simplifying the system. For the two-binding-site model, protein-DNA loading titrations ratios of 1:1, 2:1, 4:1, 6:1 and 10:1 were employed, analogous to the approach used in Zhang et al. (31). In addition, single-wavelength sedimentation velocity experiments were conducted individually on purified NanR and pure DNA using wavelengths of 280 nm and 260 nm, respectively (*SI Appendix*, Fig. S4*A*). Together, these served as the controls for the experiment, where any observation of a positive shift in sedimentation coefficient upon mixing supports the formation of a protein-DNA complex.

Following data analysis, the spectral profiles of NanR and DNA were deconvoluted into individual sedimentation coefficient distributions for each titration performed (*SI Appendix*, Fig. S4*B–F*). With increases in the protein concentration, we clearly observed co-migration relative to the individual NanR and DNA controls (*SI Appendix*, Fig. S4*A*), which is consistent with complex formation. Peak integration of these co-migrating species suggested that NanR initially binds at a 2:1 molar ratio, forming a dimeric protein-DNA hetero-complex comprising one NanR dimer bound to DNA with a sedimentation coefficient of ~4.4 S. As shown in *SI Appendix*, Fig. S4*B*, the co-migrating peaks overlaid perfectly. As the concentration of NanR was increased, we observed the formation of a broad reaction boundary. This phenomenon indicates sub-saturation of the complex, representing multiple and dynamic intermediate species that are mediated by mass action, similar to the observations made in Zhang et al. (31). As a result, peak integration at these mixing concentrations serves as an average of the species present and therefore gives rise to non-integral molar ratios. Integration of the co-migrating regions revealed molar ratios of 4:1 and 6:1, with measured sedimentation coefficients of ~6.4 S and ~ 8.6 S. These values are consistent with tetrameric and hexameric protein-DNA hetero-complexes, comprising two or three NanR dimers bound to DNA, respectively. A further increase in protein concentration led to a narrowing

of this reaction boundary towards the 6:1 molar ratio and the appearance of a new peak, corresponding to free protein. The presence of excess protein, free of any co-migrating DNA, indicated that complex formation had reached saturation. This assembly process is presented as a model in *SI Appendix*, Fig. S4*G*. In addition to characterizing the assembly process of these interacting partners, we also examined their hydrodynamic properties to provide validation of the observed complexes. Combining these properties with a weight-averaged partial specific volume for each oligomeric assembly provided a way of calculating the molar mass for each respective complex. We determined that the dimeric and tetrameric protein-DNA hetero-complexes were both in excellent agreement with their theoretical molar mass values (*SI Appendix*, Table S2). However, the formation of a hexameric complex that is consistent with three bound NanR dimers was unexpected considering we only observed two species by EMSA (*SI Appendix*, Fig. S1*C*). We hypothesize that because NanR can accommodate overlapping binding sites, the label-free oligonucleotide used in these assays may be sterically unhindered, forming protein-mediated bridges since NanR can accommodate the full recognition site by overlapping individual GGTATA repeats (6; 12).

Having characterizing the assembly process with two binding sites, we repeated the MWL-SV experiment using the full recognition site with three GGTATA repeats, employing protein-DNA titrations ratios of 1:1, 3:1, 6:1, and 10:1. Fig. 3*A*–*E* shows the deconvoluted sedimentation coefficient distributions for the titration series and the pure protein and DNA controls. Here, analogous to the previous model, we observed a clear co-migration of the NanR and DNA signals, which proceeded through several assembly steps and involved a broad reaction boundary. Initially, integration of the peak with a sedimentation coefficient of ~5.52 S gave a 2:1 molar ratio, consistent with the formation of a dimeric protein-DNA hetero-complex. In addition, integration of the peak with a sedimentation coefficient of ~7.39 S gave a 4:1 molar ratio, consistent with a tetrameric protein-DNA hetero-complex (Fig. 3*B*–*C*). Together, these two peaks were located within a broad reaction boundary, as was observed in the MWL-SV experiments using the two-binding-site model (*SI Appendix*, Fig. S4*B*–*F*). This indicated that the system was still at sub-saturation and contained multiple intermediate species mediated by mass action effects. As the concentration of NanR was increased further, the reaction boundary remained, although a shift to a higher sedimentation coefficient of ~8.27 S was observed. Integration of this peak gave a 6:1 molar ratio, consistent with a hexameric protein-DNA hetero-complex (Fig. 3*D*). Upon increasing the concentration of NanR even further, the 6:1 molar ratio remained unchanged. However, a slight increase in sedimentation coefficient was observed, and the co-migrating peak became more defined. In addition, we also observed the presence of free protein (Fig. 3*E*). Together, these features indicated that the system had reached saturation. Further analysis of the hydrodynamic properties gave measured molar mass values for the 1:1 and 6:1 molar ratio that were consistent with

the theoretical molar masses of the dimeric and hexameric protein-DNA hetero-complexes, respectively (*SI Appendix*, Table S3). Because the 4:1 molar ratio lies within the broad reaction boundary, an accurate diffusion coefficient could not be obtained. Thus, a molar mass could not be quantitively measured.

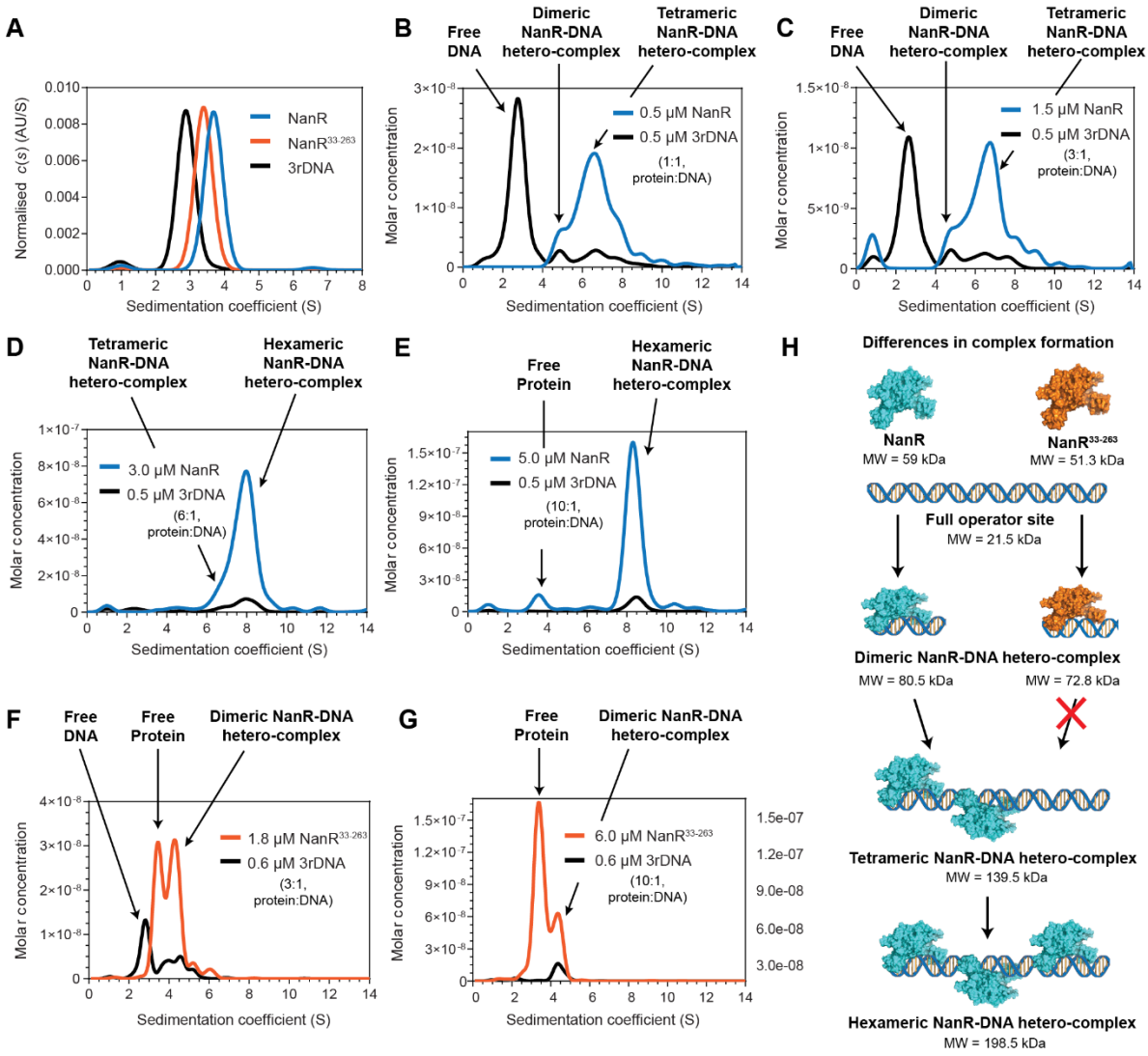


Fig. 3 | The stoichiometry of the NanR-DNA hetero-complex**.** The deconvoluted sedimentation coefficient distributions for the titration series are shown. (***A***) Separate controls of NanR (blue), NanR$^{33\text{-}263}$ (orange) and full operator site (3rDNA, black). (***B***) 0.5 μM NanR. (***C***) 1.5 μM NanR. (***D***) 3.0 μM NanR. (***E***) 5.0 μM NanR. (***F***) 0.6 μM NanR$^{33\text{-}263}$. (***G***) 6.0 μM NanR$^{33\text{-}263}$. A clear shift in the sedimentation coefficient is observed with increasing protein concentration, consistent with the formation of higher order hetero-complexes. The broad reaction boundary indicates tight binding and represents multiple, dynamic intermediate species that are mediated by mass action. This is only apparent in full-length NanR, as N-terminally truncated NanR$^{33\text{-}263}$ appears to reach saturation at the lower order (dimeric) hetero-complex. The presence of excess protein, free of any co-migrating DNA indicates that complex formation has reached saturation. All plots are presented as *g*(*s*) distributions using the molar concentration of monomeric NanR. (***H***) A model to illustrate the differences in complex assembly between NanR (blue) and NanR$^{33\text{-}263}$ (orange).

Taken together, these results indicate that we successfully characterized the stoichiometry of the NanR-DNA hetero-complex, in which NanR dimers sequentially bind DNA to ultimately form a hexameric complex bound to one DNA molecule. A model of the assembly process is presented in Fig. 3*H*. The analysis is complemented by the hydrodynamic information from each interacting partner, providing accurate molar mass measurements for the majority protein-DNA complexes across the two MWL-SV experiments (*SI Appendix*, Tables S2–3)**.** Further, we found that complex formation is concentration-dependent, supporting EMSA experiments (6; 12) (Fig. 2) and conclude that the initial assembly of the dimeric NanR-DNA hetero-complex plays an integral role in the formation of the higher-order hetero-complexes, which proceeds through a broad reaction boundary before reaching saturation. This reaction boundary is indicative of a reversibly associating system with multiple intermediates, as opposed to the presence of discrete, non-interacting species. We note that DNA was always present in sub-stoichiometric amounts, and therefore was consumed by NanR binding as dimers to the DNA, hence free DNA was never observed in these experiments. This suggests that the K_D for this assembly process is very low, supporting the nanomolar affinity defined in this study (Fig. 2).

N-terminally truncated NanR derivatives bind DNA with significantly reduced affinity.

Cooperative binding of transcriptional regulators to DNA is typically driven by protein-protein interactions between neighboring protomers (32)—a process critical to gene regulation. Our EMSA and analytical ultracentrifugation binding experiments both demonstrated that NanR exhibits positive cooperative behavior, as shown by the sigmoidal dependency of the binding isotherm (Fig. 2*C*). To investigate what drives this cooperativity, we performed a sequence homology search within the PDB. We identified that NanR possesses a predominantly disordered 32-residue N-terminal extension of the DNA-binding domain (*SI Appendix*, Fig. S5), which is considerably longer than those of similar structures in closely related GntR transcriptional regulators. Notably, this extension contains four serine residues (at positions 9, 14, 15 and 24), four arginine residues (at positions 20, 23, 25 and 29), and 11 hydrophobic residues spread throughout. To assess the role of this extension in cooperativity, we deleted the 32 N-terminal residues to generate a truncated version of NanR, designated $NanR^{33\text{-}263}$. Using MWL-SV experiments with the full recognition site and protein-DNA mixing titrations of 3:1 and 10:1, we observed a significant change in the assembly process (Fig. 3*F–G*) compared with that of the full-length protein (Fig. 3*B–E*). Following peak integration of the co-migrating regions, we determined that the N-terminally truncated NanR bound DNA with a 2:1 molar ratio and a sedimentation coefficient of ~4.39 S, which is consistent with formation of a dimeric protein-DNA hetero-complex and values observed with that of the full-length NanR. However, we also observed the presence of free protein (Fig. 3*F*), unlike the full-length NanR at this titration, which therefore indicates weaker binding. Upon increasing concentration of NanR, the co-migrating peak for the dimeric protein-DNA hetero-complex

became more defined with a sedimentation coefficient of ~4.46 S, and the signal corresponding to free DNA disappeared, indicating that the system had reached saturation. The measured molar mass of this dimeric protein-DNA hetero-complex was in excellent agreement with its theoretical molar mass (*SI Appendix*, Table S4), providing further confidence in the 2:1 protein-DNA molar ratio. Interestingly, while we showed that full-length NanR bound DNA and ultimately formed a hexameric protein-DNA hetero-complex, N-terminally truncated NanR was only capable of binding as a single protomer, forming a dimeric protein-DNA hetero-complex at the titration range tested.

Therefore, removal of the 32-residue N-terminal extension appears to perturb the ability of the repressor to form higher-order hetero-complexes *via* a concentration-dependent assembly. Consequently, we propose that this N-terminal extension is the driving force of the positive cooperative behavior, facilitated by protein-protein interaction. We also hypothesize that the initial binding to DNA causes a conformational change of NanR that facilitates a favorable interaction with neighboring protomers through a cooperative process mediated by the N-terminal extension. As this extension is significantly larger than those found in closely related GntR members (*SI Appendix*, Fig. S5), we believe this cooperative mechanism is unique to *E. coli* NanR, among characterized GntR family members, and is required to maintain tight, coordinated control of the sialoregulon.

The structure of the NanR Neu5Ac complex reveals a metal-binding site.

To gain insight into the molecular mechanism of regulation, we solved the structure of *E. coli* NanR in complex with Neu5Ac and a zinc ion (Fig. 4*A*–*B*). This intrinsically bound zinc ion was identified through inductively coupled plasma mass spectrometry and X-ray absorption spectroscopy, then used to phase the structure using a combined single-wavelength anomalous dispersion and molecular replacement strategy to a maximum resolution of 2.1 Å (*SI Appendix*, Table S5). This discovery places NanR in a small and distinct family of GntR transcriptional regulators, including TM0439, that contain a metal-binding site (33; 34).

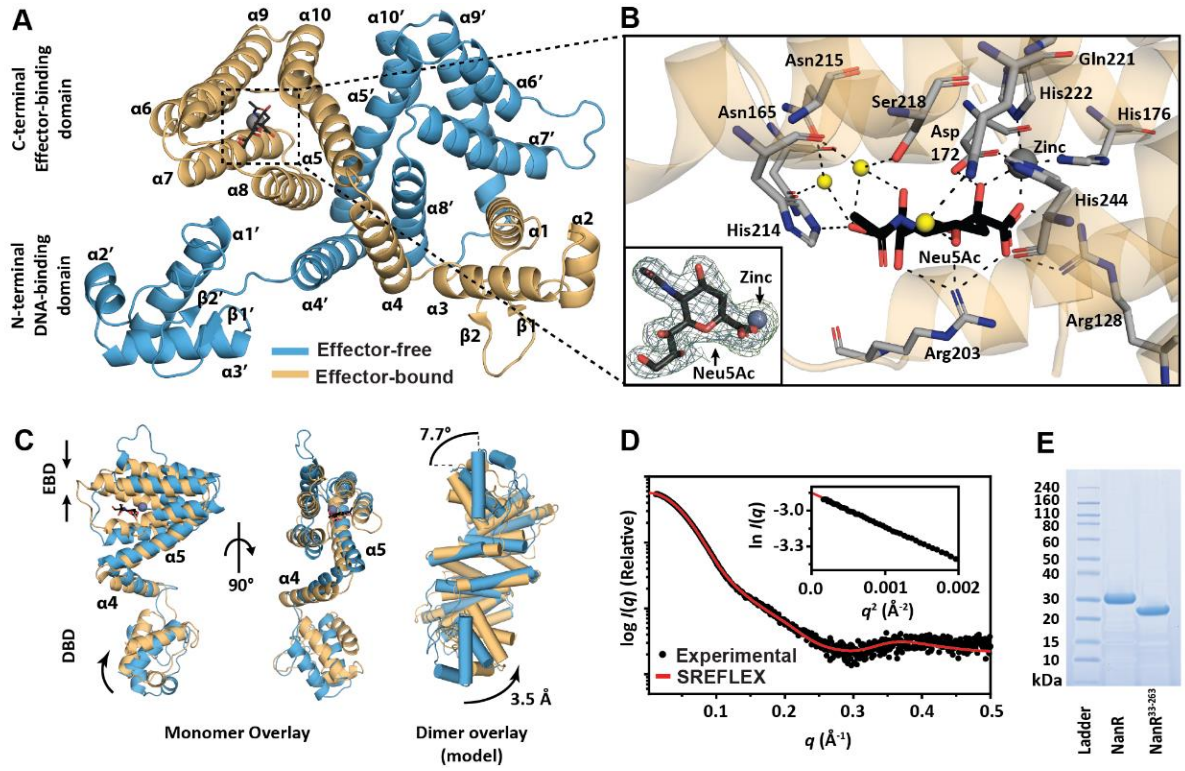


Fig. 4 | Overall structure of *E. coli* NanR in complex with Neu5Ac and zinc. (***A***) Cartoon representation of the domain-swapped dimeric structure. NanR comprises an N-terminal DNA-binding domain, with the characteristic winged helix-turn-helix motif, and a C-terminal effector binding domain, analogous to other GntR members. The effector-bound and the effector-free are shown in beige and cyan, respectively, and all secondary structure elements are labelled. Prime denotes the effector-free monomer. (***B***) The effector-binding site is located within the large polar cavity of the C-terminal domain. The direct or water-mediated (black dashes) hydrogen bonding residues (grey sticks) that coordinate Neu5Ac in its β-anomeric form and support zinc in an octahedral geometry are indicated, while water molecules are depicted as yellow spheres. The omit map corresponding to Neu5Ac (black sticks) and zinc (grey sphere) is shown as an insert. Here, the 2F*o*-F*c* electron density map (1.0 σ, blue mesh) and the mF*o*-F*c* omit electron density map (3.0 σ, green mesh) is shown. (***C***) An overlay of the effector-bound monomer (beige) and effector-free monomer (cyan) illustrates effector-induced conformational changes, including: 1) a restriction of the all-α-helical bundle of the effector-binding domain (EBD); and 2) a hinging motion between α5 and α4 causing a displacement of the DNA-binding domain (DBD) by 7.7° and 3.5 Å. Superimposition of a dimeric model for each distinct state, depicting helices as cylinders, further illustrates these changes. (***D***) The experimental scattering profile for NanR$^{33–263}$ (black dots) is consistent with the theoretical scattering of the atomic structure (red line, X^2 value of 0.32). The Guinier plot shown as an insert, highlights linearity at low *q*, providing confidence that the data is free of significant aggregation or inter-particle interference. (***E***) SDS-PAGE analysis showing protein molecular weight ladder (Lane 1), and purified *E. coli* NanR (Lane 2) and NanR$^{33–263}$ (Lane 3) following size exclusion chromatography.

Analogous to other GntR members of the superfamily, NanR has a two-domain architecture comprising an N-terminal DNA-binding domain (α1–α3, β1–β2), and an all α-helical C-terminal effector-binding domain (α4 to α10) (Fig. 4*A*). The antiparallel two-stranded β-sheet defines the "wing" of the characteristic wHTH motif—a trademark of the GntR family (18). The presence of seven α-helices in the C-terminal domain identifies NanR as member of the FadR subfamily (17). Helices α5–α10 are arranged into an antiparallel bundle and play a role in both effector binding and dimerization, while helix α4 serves as a flexible linker connecting the N-terminal domain to the all α-helical bundle of the C-terminal domain. Moreover, this linker is believed to play a role in the allosteric mechanism (18). Our structure of NanR contains two monomers (Chain A, Lys31–Asn246; Chain B, Lys31–His244) of NanR in the asymmetric unit, which is consistent with the observation of a dimer in solution (*SI Appendix*, Fig. S2). Remarkably, while we were unsuccessful in collecting diffraction data suitable for determining the apo structure, the structure presented in the current study captures the effector-bound conformation in one monomer (Chain A), while the opposing monomer (Chain B) is effector-free. Superimposition of these asymmetric monomers revealed a root-mean-square deviation (RMSD) of 4.55 Å over 207 equivalent Cα atoms, indicating that a conformational change had occurred between monomers (Fig. 4*C*). The 30 N-terminal residues, which define the N-terminal domain extension (*SI Appendix*, Fig. S5), the 17 C-terminal residues in chain A and the 19 C-terminal residues in chain B were not included in the final model because we did not observe any corresponding electron density. These regions, which we believe to be highly flexible were likely cleaved during incubation with chymotrypsin during crystallization, which was essential for structure determination.

Interestingly, the dimeric assembly of NanR showed a domain-swapped architecture. Mediated by the flexible α4 helix, this was achieved through an exchange of the N-terminal domain between monomers (Fig. 4*A*). This feature is conserved in FadR and LldR, which are involved in the regulation of fatty acids (35) and L-lactate (34), respectively. Using the PDBePISA server (36), we determined that approximately 2004 $Å^2$ of accessible surface area was buried upon dimerization, while the total estimated ΔG^{int} value of the dimer interface was approximately −25.9 kcal mol^{-1}, corresponding to 15% of the total surface area of each monomer. Primarily, this dimer interface was formed *via* the helical bundle of each C-terminal domain, facilitated by a hydrophobic core along helices α4, α5, and α8. However, polar and salt bridge interactions helped stabilize the dimer interface, most notably between the N- and C-terminal domains of opposing monomers. Interestingly, we discovered that polyethylene glycol (PEG) was buried in this dimer interface. We initially hypothesized that this electron density fitted a bacterial phospholipid. However, following lipid extraction of the purified protein and subsequent quadrupole time-of-flight mass spectrometry, no lipid signals were detected. Thus, considering PEG was present in

the crystallization buffer, it was modelled into the density. PEG above 5 kDa has been reported to exhibit more amphiphilic characteristics, permitting such an interaction (37).

To verify that the dimeric assembly in the crystal structure is biologically relevant in solution, we conducted SAXS experiments with the N-terminally truncated NanR construct. As shown in Fig. 4*D*, the theoretical scattering profile generated from the crystal structure of NanR was in excellent agreement with the experimental scattering profile of $NanR^{33-263}$, with a reduced X^2 value of 0.32, suggesting that the dimeric assembly observed in the crystal structure is shared globally in solution. This fit was facilitated using SREFLEX, which enabled flexible refinement of the atomic model to accommodate conformational rearrangement in solution (38). In addition, Kratky analysis demonstrated that NanR is flexible in solution, as although a bell-shaped curve was observed, the plot did not converge to the *q* axis (*SI Appendix*, Fig. S3*B*). This feature is likely associated with the flexible linker between the N- and C-terminal domains and provides justification for the higher overall *B*-factor observed for NanR (*SI Appendix*, Table S5).

Insight into the effector-binding allosteric mechanism.

The C-terminal domain has a large polar cavity buried within the all-α-helical bundle that forms the effector-binding site. In the current study, electron density of Neu5Ac was unambiguously observed in its β-anomeric form (Fig. 4*B*). This observed stereochemistry is consistent with reports that bacteria who scavenge host-derived Neu5Ac in the α-anomeric form from glycoconjugates possess a mutarotase that rapidly converts the amino sugar to the more thermodynamically stable β-anomer (11) (Fig. 1*A*). However, this evidence contradicts the hypothesis that the α-anomer of Neu5Ac is the physiological inducer of the sialocatabolic system (6).

The stereochemistry is defined by Arg128, which forms a salt bridge with the negatively-charged carboxylate group of Neu5Ac (Fig. 4*B*). The presence of a basic residue within the effector-binding site is a common feature amongst repressors who preferentially bind negatively-charged effector molecules (25), as well as the sialic acid-binding protein SiaP, found in the periplasm (39). Further coordination was predominantly provided by direct or water-mediated hydrogen bonding involving Asn165, Asp172, His176, Arg203, His214, Asn215, Ser218, Gln221, and His244, as well as hydrophobic interactions with Phe168, Ile200, and Leu245 (Fig. 4*B* and *SI Appendix*, Fig. S6). In addition to Neu5Ac, the zinc ion was also buried within the cavity through an octahedral geometry with a refined *B*-factor of 36.7 $Å^2$, which is consistent with a bound metal ion. Specifically, this coordination involved OD_1 of Asp172, NE_2 of His176, His222, and His244, and the O_1 carboxyl and O_2 hydroxyl moieties of Neu5Ac (Fig. 4*B* and *SI Appendix*, Fig. S6). In each case, the bond distance was in the range of 2.0–2.2 Å. Collectively, these

residues are part of a conserved fingerprint motif (Arg-X_3-Glu-X_{40}-Asp-X_4-His-$X_{\sim 50}$-His-$X_{\sim 20}$-His) that is present in GntR transcriptional regulators that bind metal ions, where the histidine residues form a triangular, propeller-like architecture (33). To ensure these histidine residues can coordinate the zinc ion *via* their NE_2 atoms, residues Glu132 and Gln221 stabilize the $N\delta_1$ protonated tautomer through hydrogen bonds of 2.6 and 2.9 Å, respectively.

As mentioned above, our structure captures NanR in both effector-bound and effector-free (apo) states. This discovery allows us to probe the conformational changes that occur following effector binding. We found that effector binding triggers a conformational change, compacting the all-α-helical bundle, which in turn causes an Arg128-mediated shift of helix α5 towards the helical linker, α4. This propagation of the allosteric signal results in the displacement of the DNA-binding domain by 7.7° over 3.5 Å, which we believe would disrupt DNA binding. To illustrate this conformational change, we constructed a model of the dimeric effector-free and effector-bound structures (Fig. 4*C*). We observed that the displacement of the N-terminal DNA-binding domain occurred in opposing directions, supporting our hypothesis that this change would decrease the binding affinity for DNA. The conformational change was further supported by our solution studies where we observed a small change in shape (*SI Appendix*, Fig. S1 & Table S1). However, EMSA experiments did not show a significant change in affinity in the presence of Neu5Ac and zinc (*SI Appendix*, Fig. S1*D*–*E*). This suggests that perhaps multiple ligands are required to bind the effector or a complex interplay between additional proteins is required. .

Conformation of DNA-bound NanR.

Next, we conducted single-particle cryo-electron microscopy (cryo-EM) experiments to investigate the DNA-bound conformation of NanR. Experiments were performed using a protein-DNA ratio titration of 2:1 after the observation of a stable dimeric protein-DNA hetero-complex, in solution, comprising one NanR dimer bound to DNA during MWL-SV experiments with the two-binding-site model (*SI Appendix*, Fig. S4*C*). To further reduce any sample heterogeneity and remove any aggregates, the complex was purified using size-exclusion chromatography (*SI Appendix*, Fig. S7*A*). Once imaged, micrographs revealed an even particle distribution (Fig. 5*A*), while two-dimensional (2D) class averages showed defined secondary structural elements and the presence of DNA, supporting complex formation (Fig. 5*B*). Despite the size of the dimeric NanR-DNA hetero-complex (70.5 kDa), a volta phase plate was not utilized in this experiment. Following 3D classification, ~140K particles were used for 3D refinement, producing a 3.9 Å resolution electron scattering potential map, as determined by the 'gold standard' Fourier shell correlation criterion (=0.143) (Fig. 5*C*).Using the crystal structure of NanR as a reference, we constructed an atomic model of the dimeric NanR-DNA hetero-complex, fit the model to the cryo-EM map using PHENIX, and refined the model using atomic restraints (40) (Fig. 5*D*).

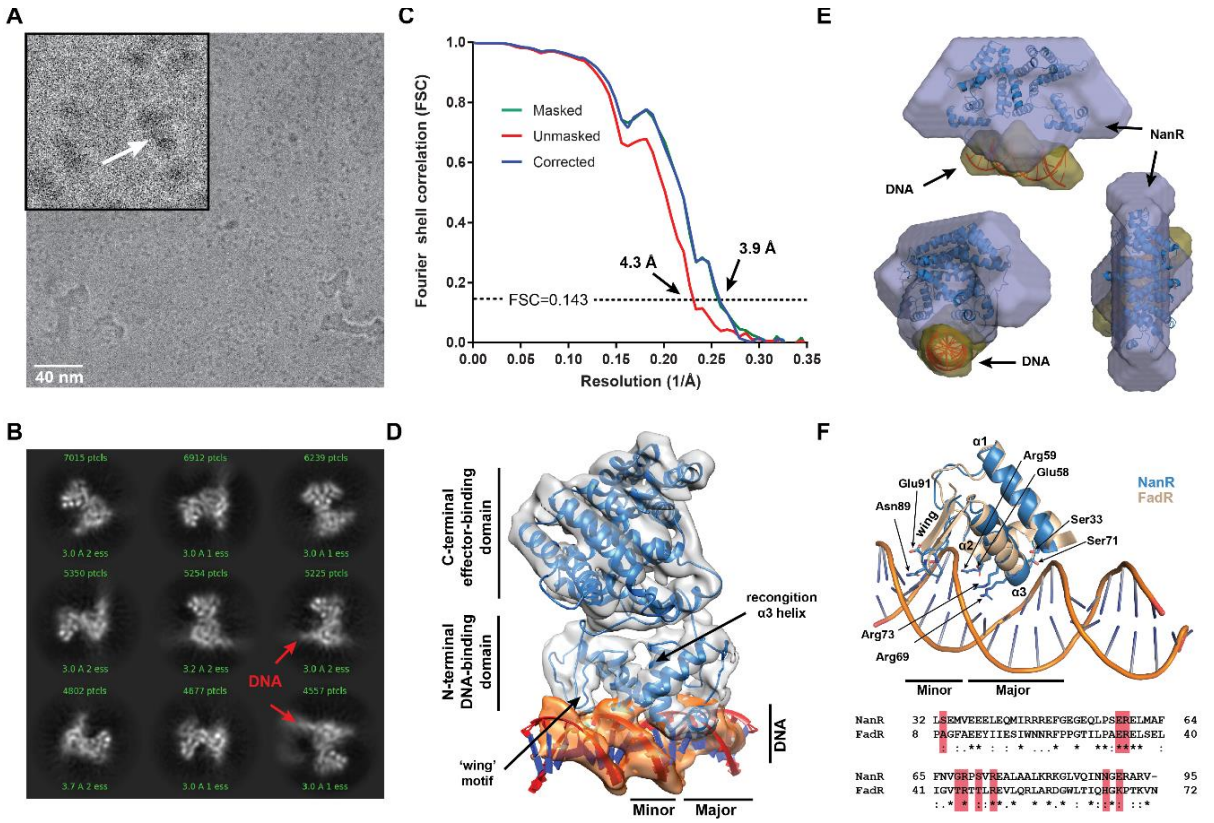


Fig. 5 | The DNA-bound conformation of *E. coli* NanR. (***A***) Electron micrograph of the dimeric NanR-DNA complex (scale bar, 40 nm). Insert: complex particle (white arrow). (***B***) The 2D class averages of the dimeric protein-DNA hetero-complex depict secondary structure elements and the presence of DNA in projection. (***C***) Fourier shell correlation plots indicate an estimated resolution of 3.9 Å for the masked (green line) and corrected (blue line) initial 3D map, and 4.3 Å for the unmasked (red) initial 3D map. (***D***) 3D electron scattering potential map and docked model of the dimeric NanR-DNA hetero-complex. The N-terminal DNA-binding domains (DBD), the C-terminal effector binding domains (EBD) and density are consistent with the N-terminal extension of NanR, perpendicular to the DNA. (***E***) Multiphase *ab initio* model reconstructed from the scattering profiles of NanR, DNA and the dimeric NanR-DNA hetero-complex in solution. Despite its low resolution, this model supports the binding orientation observed in the electron scattering potential map. (***F***) A model of the NanR N-terminal DNA binding domain (blue) superimposed on the structure of the FadR-DNA complex [PDB ID: 1HW2, in beige]. For clarity, only one N-terminal domain is shown. The wing is shown to bind in the minor groove, whereas the HTH motif binds in the major groove. The NanR amino acid residues predicted to be involved in DNA binding based on sequence conservation are shown as sticks and labelled. The sequence alignment for this region is shown in the lower panel with predicted residues highlighted (red).

The 2D class averages (Fig. 5*B*) and refined 3D model, docked to the electron scattering potential map (Fig. 5*D*) present the DNA-bound conformation for *E. coli* NanR. We observed that NanR binds DNA between the N-terminal DNA-binding domains, making contact in the major groove with the recognition α3 helix (Fig. 5*D*). The distance between the N-terminal DNA-binding domains was ~20 Å, forming a positively-charged DNA-binding cleft to accommodate the diameter of the DNA double helix (~18 Å) (*SI Appendix*, Fig. S8). In the absence of DNA, these domains were separated by ~30 Å (Fig. 4*A*), This conformation change, between each state, was facilitated by the flexible linker, helix α4.

To verify the orientation of NanR and DNA in solution, we collected SAXS data for the dimeric NanR-DNA hetero-complex (*SI Appendix*, Fig. S3*C*) and reconstructed a multiphase *ab initio* model using MONSA (41) (Fig. 5*E*). This complex was assembled in an identical manner to the single-particle cryo-EM experiments using the two-binding-site model and NanR. Using SUPCOMB (42), the atomic structure of NanR and an atomic model of the DNA oligonucleotide were superimposed with the multiphase model with normalized spatial discrepancies of 3.3 and 1.6, respectively. Although low resolution, this *ab initio* model supported the binding conformation observed in the cryo-EM map, where the C-terminal domain of NanR is primarily orientated in parallel with the helical axis of DNA.

Together, the 3D cryo-EM model (Fig. 5*D*) and the multiphase *ab initio* model (Fig. 5*E*) present the DNA-bound conformation of the dimeric protein-DNA hetero-complex. However, the resolution of the structural data restricted our ability to elucidate the specific amino acid residues that are responsible for DNA binding. Therefore, we superimposed the N-terminal DNA-binding domain of NanR onto the structural coordinates of the GntR-type transcriptional regulator FadR-DNA hetero-complex (PDB ID – 1HW2 (43)) to generate a NanR-DNA model (Fig. 5*F*). The results showed that, the N-terminal domain of NanR and FadR aligned well, with an RMSD of 0.984 Å, with the recognition α3 helix bound within the major groove, and the wing motif interacting with the minor groove of DNA. This was consistent with our cryo-EM model (Fig. 5*D*). However, as the recognition sites vary between NanR and FadR in both direction and length, this analysis only provides a foundation to examine the role of conserved residues that are likely to play a functional role in DNA-binding.

Based on sequence comparisons (Fig. 5*F*), nine putative DNA-binding amino acid residues were identified. Firstly, residue Ser31 in helix α1, residue Glu58 in helix α2, residues Gly68 and Ser71 of helix α3, and residue Glu91 in the wing motif may form hydrogen bonds with the phosphate backbone. Secondly, residue Arg59 in helix α2, residues Arg69 and Arg73 in helix α3, and residue Asn89 in the wing motif are likely to make sequence-specific contacts with one of the DNA bases within the recognition sequence (Fig. 1*D*). Furthermore, these four residues that are implicated in specific base contacts all

map to the most positive electrostatic surface of the N-terminal domain (*SI Appendix*, Fig. S8), and hence support a role in DNA binding within *E. coli* NanR.

Structural insight into multimeric protein-DNA hetero-complex assembly.

Having characterised the stoichiometry of the NanR-DNA complex and demonstrated that NanR ultimately forms a hexameric protein-DNA hetero-complex, comprising three NanR dimers bound to DNA, we conducted further single-particle cryo-EM experiments to investigate this multimeric assembly process. Because a stable hexameric protein-DNA hetero-complex was observed during MWL-SV experiments, further single particle cryo-EM imaging was performed using a protein-DNA titration ratio of 10:1 with the full-length operator site (Fig. 3). The hexameric protein-DNA hetero-complex was purified using size-exclusion chromatography (*SI Appendix*, Fig. S7*B*). Once imaged, 2D class averages clearly identified the helical axis of DNA and revealed the presence of multiple NanR protomers assembling along this oligonucleotide (Fig. 6). Further classification and interpretation allowed us to construct a 3D model for the multimeric protein-DNA hetero-complex assembly process using our refined cryo-EM structure of the dimeric protein-DNA hetero-complex (Fig. 5*D*). As observed previously, NanR was primarily oriented parallel to the helical axis of DNA (Fig. 6). Upon formation of the higher-order complexes, each additional NanR dimer appeared to bind very closely to each other. This observation is consistent with NanR accommodating the full recognition site by overlapping individual GGTATA repeats (6). In addition, the proximity of NanR supports the involvement of protein-protein interactions in providing stabilization or in anchoring the complex together.

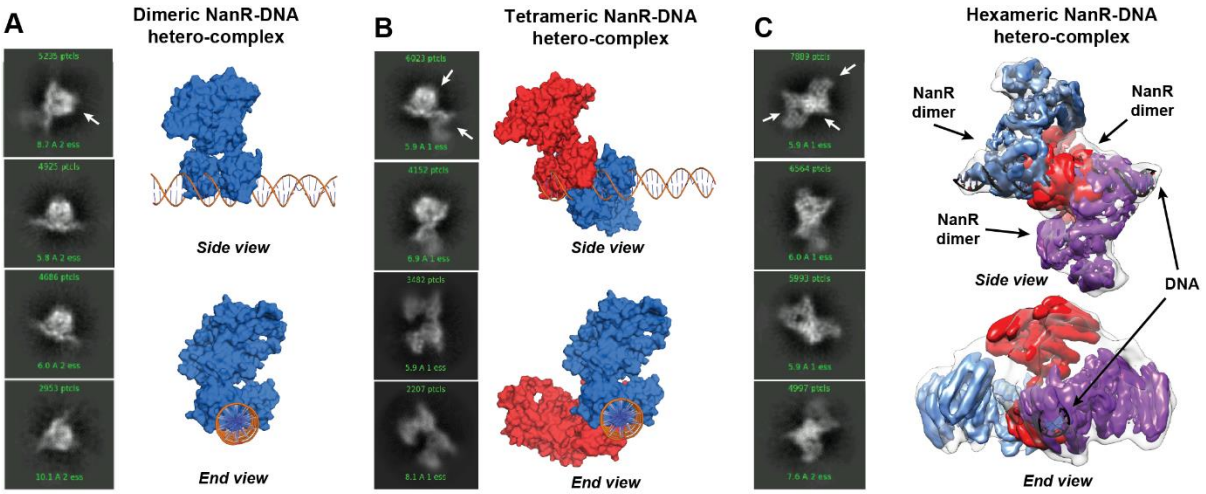


Fig. 6 | Structural insight into multimeric NanR-DNA hetero-complex assembly. (***A***) 2D class averages showing a single NanR dimer bound to DNA in projection (left).and a model of the dimeric NanR-DNA hetero-complex using these 2D classes (right). (***B***) 2D class averages showing two NanR dimers bound to DNA in projection (left) and a model of the tetrameric NanR-DNA hetero-complex (right). (***C***) 2D class averages showing three NanR dimers bound to DNA in projection (left) and a Cryo-EM structure of the hexameric NanR-DNA hetero-complex (right). All models are shown as surface representations using the DNA-bound cryo-EM structure (Fig. 6*D*). Together, these serve as a structural model for the multimeric NanR-DNA complex assembly process.

Conclusions.

In this study, we report the biophysical and structural characterization of the GntR-type transcriptional regulator NanR from *E. coli*. Our binding studies demonstrate NanR exhibits nanomolar affinity and binds DNA as sequential dimers, ultimately saturating at a hexameric protein-DNA hetero-complex with three NanR dimers coordinating DNA. The MWL-SV data provide an excellent case study for analysis of protein-DNA interactions. In addition, we think it likely that the potential of this emerging approach could be expanded to other complex systems, provided that the spectral properties of the interacting partners can be exploited (29). Using full-length NanR we define this assembly process is facilitated by positive cooperativity. Removal of a 32-residue, N-terminal extension to the DNA-binding domain perturbs this cooperative binding, suggesting this feature, is the driving force for this behavior. We hypothesize that this extension plays a fundamental role in both the recruitment of additional NanR dimers and towards anchoring the NanR dimers to maintain a high-order hetero-complex, likely through protein-protein interactions. Using crystallography, we solved the structure of NanR in complex with Neu5Ac and zinc, which provided the first evidence of a metal ion playing a direct role in the binding of an effector molecule, expanding our understanding of the metal-binding transcriptional regulators of the GntR family. This structure also illustrated the conformational changes that occur as part of the

allosteric mechanism to reorient the N-terminal DNA-binding domains. This provided novel insight into the allosteric mechanism to attenuate gene expression and strong evidence to support the suggestion that Neu5Ac is the effector molecule *in vivo*. In addition, we report the structure of the dimeric protein-DNA hetero-complex by single particle cryo-EM, and structurally verified the multimeric protein-DNA hetero-complex assembly process that was observed in our MWL-SV experiments.

Materials and Methods

All reagents, media and consumables were purchased from Thermo Fisher Scientific and Sigma-Aldrich, unless otherwise stated. A detailed description of the protein cloning, expression and purification of *E. coli* NanR, double-stranded DNA preparation, SAXS analysis, single- and multi-wavelength sedimentation velocity analysis, EMSA assays, determination of the K_D, protein crystallization, phase determination and X-ray structure refinement, and single-particle cryo-EM sample preparation, data acquisition and data processing are described in *SI Appendix*. X-ray coordinates have been deposited in the PDB under the ID code 60N4.

Acknowledgements

This work was supported by the following funding to R.C.J.D: 1) the New Zealand Royal Society Marsden Fund (contract UOC1506); 2) a Ministry of Business, Innovation and Employment Smart Ideas grant (contract UOCX1706); and 3) the Biomolecular Interaction Centre (University of Canterbury), and also supported by the following funding to C.R.H: 1) the Canterbury Medical Research Foundation, and 2) the Maurice Wilkins Centre. We acknowledge the Australian Synchrotron, the staff at the MX2 beamline and Dr Nigel Kirby and Dr Tim Ryan for provision of synchrotron facilities and for their assistance in using the beamlines. We would also like to acknowledge Dr Yee-Foong Mok for his assistance with fluorescence-detection sedimentation velocity experiments at the Macromolecular Interactions Facility, Bio21 Institute, University of Melbourne, Dr Trushar Patel (University of Lethbridge) for his assistance with SAXS data analysis, Dr Janet Newman (Collaborative Crystallisation Centre) for her assistance during protein crystal optimization, and Dr Tamsin Sheen for proofing this work. MWL-SV experiments were performed at the Canadian Center for Hydrodynamics, University of Lethbridge. B.D acknowledges support from grants NIH-GM120600 and NSF-ACI-1339649. Supercomputer calculations were performed on Comet at the San Diego Supercomputing Center, and Jetstream and Stampede at the Texas Advanced Computing Center (supported through NSF/XSEDE grant TG-MCB070039N to B.D), and on Lonestar-5 at the Texas Advanced Computing Center (supported through UT grant TG457201 to B.D).

References

1. Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R., & Gordon, J. I. (2008). Evolution of mammals and their gut microbes. *Science, 320*, 1647-1651.
2. Jeong, H. G., Oh, M. H., Kim, B. S., Lee, M. Y., Han, H. J., & Choi, S. H. (2009). The capability of catabolic utilisation of *N*-acetylneuraminic acid, a sialic acid, is essential for *Vibrio vulnificus* pathogenesis. *Infection & Immunity, 77*, 3209-3217.
3. Chang, D. E., Smalley, D. J., Tucker, D. L., Leatham, M. P., Norris, W. E., Stevenson, S. J., Anderson, A. B., Grissom, J. E., Laux, D. C., Cohen, P. S., & Conway, T. (2004). Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 7427-7432.
4. Gorke, B., & Stulke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Reviews Microbiology, 6*, 613-624.
5. Deutscher, J. (2008). The mechanisms of carbon catabolite repression in bacteria. *Current Opinion in Microbiology, 11*, 87-93.
6. Kalivoda, K. A., Steenbergen, S. M., Vimr, E. R., & Plumbridge, J. (2003). Regulation of sialic acid catabolism by the DNA binding protein NanR in *Escherichia coli*. *Journal of Bacteriology, 185*, 4806-4815.
7. Baos, S. C., Phillips, D. B., Wildling, L., McMaster, T. J., & Berry, M. (2012). Distribution of sialic acids on mucins and gels: A defense mechanism. *Biophysical Journal, 102*, 176-184.
8. Vimr, E. R., Kalivoda, K. A., Deszo, E. L., & Steenbergen, S. M. (2004). Diversity of microbial sialic acid metabolism. *Microbiol. Molecular Biology Reviews, 68*, 132-153.
9. Angata, T., & Varki, A. (2002). Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. *Chemical Reviews, 102*, 439-469.
10. Almagro-Moreno, S., & Boyd, E. F. (2009). Insights into the evolution of sialic acid catabolism among bacteria. *BMC Evolutionary Bioogy, 9*, 118.
11. Vimr, E. R. (2013). Unified theory of bacterial sialometabolism: how and why bacteria metabolise host sialic acids. *International Scholarly Research Notices Microbiology, 2013*, 816713.
12. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
13. Condemine, G., Berrier, C., Plumbridge, J., & Ghazi, A. (2005). Function and expression of an *N*-acetylneuraminic acid-inducible outer membrane channel in *Escherichia coli*. *Journal of Bacteriology, 187*, 1959-1965.
14. Steenbergen, S. M., Jirik, J. L., & Vimr, E. R. (2009). YjhS (NanS) Is Required for *Escherichia coli* To Grow on 9-O-Acetylated *N*-Acetylneuraminic Acid. *Journal of Bacteriology, 191*, 7134-7139.
15. Severi, E., Müller, A., Potts, J. R., Leech, A., Williamson, D., Wilson, K. S., & Thomas, G. H. (2008). Sialic acid mutarotation is catalyzed by the *Escherichia coli* β-propeller protein YjhT. *Journal of Biological Chemistry, 283*, 4841-4849.
16. Suvorova, I. A., Korostelev, Y. D., & Gelfand, M. S. (2015). GntR Family of Bacterial Transcription Factors and Their DNA Binding Motifs: Structure, Positioning and Co-Evolution. *Public Library of Science One, 10*, e0132618-e0132618.
17. Rigali, S., Derouaux, A., Giannotta, F., & Dusart, J. (2002). Subdivision of the Helix-Turn-Helix GntR Family of Bacterial Regulators in the FadR, HutC, MocR, and YtrA Subfamilies. *Journal of Biological Chemistry, 277*, 12507-12515.
18. Jain, D. (2015). Allosteric control of transcription in GntR family of transcription regulators: A structural overview. *IUBMB Life, 67*, 556-563.
19. Ptashne, M. (2004). *A genetic switch : phage lambda revisited* (3rd ed.). Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
20. Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research, 14*, 1188-1190.
21. Brenowitz, M., Mandal, N., Pickar, A., Jamison, E., & Adhya, S. (1991). DNA-binding properties of a *lac* repressor mutant incapable of forming tetramers. *Journal of Biological Chemistry, 266*, 1281-1288.
22. Mota, L. J., Sarmento, L. M., & de Sa-Nogueira, I. (2001). Control of the arabinose regulon in *Bacillus subtilis* by AraR *in vivo*: crucial roles of operators, cooperativity, and DNA looping. *Journal of Bacteriology, 183*, 4190-4201.

23. Blancato, V. S., Repizo, G. D., Suarez, C. A., & Magni, C. (2008). Transcriptional regulation of the citrate gene cluster of *Enterococcus faecalis* Involves the GntR family transcriptional activator CitO. *Journal of Bacteriology, 190*, 7419-7430.
24. Gebhard, S., Busby, J. N., Fritz, G., Moreland, N. J., Cook, G. M., Lott, J. S., Baker, E. N., & Money, V. A. (2014). Crystal structure of PhnF, a GntR-family transcriptional regulator of phosphate transport in *Mycobacterium smegmatis*. *Journal of Bacteriology, 196*, 3472-3481.
25. Fillenberg, S. B., Grau, F. C., Seidel, G., & Muller, Y. A. (2015). Structural insight into operator dre-sites recognition and effector binding in the GntR/HutC transcription regulator NagR. *Nucleic Acids Research, 43*, 1283-1296.
26. Kataoka, M., Tanaka, T., Kohno, T., & Kajiyama, Y. (2008). The carboxyl-terminal domain of TraR, a *Streptomyces* HutC family repressor, functions in oligomerization. *Journal of Bacteriology, 190*, 7164-7169.
27. Okuda, K., Ito, T., Goto, M., Takenaka, T., Hemmi, H., & Yoshimura, T. (2015). Domain characterization of *Bacillus subtilis* GabR, a pyridoxal 5'-phosphate-dependent transcriptional regulator. *Journal of Biochemistry, 158*, 225-234.
28. Fischer, H., Neto, M. D., Napolitano, H. B., Polikarpov, I., & Craievich, A. F. (2010). Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *Journal of Applied Crystallography, 43*, 101-109.
29. Gorbet, G. E., Pearson, J. Z., Demeler, A. K., Colfen, H., & Demeler, B. (2015). Next-Generation AUC: Analysis of Multiwavelength Analytical Ultracentrifugation Data. In J. L. Cole (Ed.), *Analytical Ultracentrifugation* (Vol. 562, pp. 27-47).
30. Lawson, C. L., & Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs, NJ., Prentice-Hall, Inc.
31. Zhang, J., Pearson, J. Z., Gorbet, G. E., Colfen, H., Germann, M. W., Brinton, M. A., & Demeler, B. (2017). Spectral and Hydrodynamic Analysis of West Nile Virus RNA-Protein Interactions by Multiwavelength Sedimentation Velocity in the Analytical Ultracentrifuge. *Analytical Chemistry, 89*, 862-870.
32. Wolberger, C. (1999). Multiprotein-DNA complexes in transcriptional regulation. *Annual Reviews in Biophysical Biomolecular Structure, 28*, 29-56.
33. Zheng, M., Cooper, D. R., Grossoehme, N. E., Yu, M., Hung, L. W., Cieslik, M., Derewenda, U., Lesley, S. A., Wilson, I. A., Giedroc, D. P., & Derewenda, Z. S. (2009). Structure of *Thermotoga maritima* TM0439: implications for the mechanism of bacterial GntR transcription regulators with Zn2+-binding FCD domains. *Acta Crystallographica, Section D, Biological Crystallography, 65*, 356-365.
34. Gao, Y. G., Suzuki, H., Itou, H., Zhou, Y., Tanaka, Y., Wachi, M., Watanabe, N., Tanaka, I., & Yao, M. (2008). Structural and functional characterization of the LldR from *Corynebacterium glutamicum*: a transcriptional repressor involved in L-lactate and sugar utilization. *Nucleic Acids Research, 36*, 7110-7123.
35. van Aalten, D. M., DiRusso, C. C., Knudsen, J., & Wierenga, R. K. (2000). Crystal structure of FadR, a fatty acid-responsive transcription factor with a novel acyl coenzyme A-binding fold. *EMBO Journal, 19*, 5167-5177.
36. Krissinel, E., & Henrick, K. (2005). Detection of protein assemblies in crystals. In M. R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, & I. Fischer (Eds.), *Computational Life Sciences, Proceedings* (Vol. 3695, pp. 163-174).
37. Wu, J., Zhao, C., Lin, W., Hu, R., Wang, Q., Chen, H., Li, L., Chen, S., & Zheng, J. (2014). Binding characteristics between polyethylene glycol (PEG) and proteins in aqueous solution. *Journal of Material Chemistry, 2*, 2983-2992.
38. Panjkovich, A., & Svergun, D. I. (2016). Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis. *Physical Chemistry Chemical Physics, 18*, 5707-5719.
39. Rosa, L. T., Bianconi, M. E., Thomas, G. H., & Kelly, D. J. (2018). Tripartite ATP-Independent Periplasmic (TRAP) Transporters and Tripartite Tricarboxylate Transporters (TTT): From Uptake to Pathogenicity. *Frontiers in Cellular and Infection Microbiology, 8*, 33.
40. Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., & Adams, P. D. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica, Section D, Biological Crystallography, 68*, 352-367.
41. Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biochemical Journal, 76*, 2879-2886.
42. Kozin, M. B., & Svergun, D. I. (2001). Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography, 34*, 33-41.
43. Xu, Y., Heath, R. J., Li, Z., Rock, C. O., & White, S. W. (2001). The FadR.DNA complex. Transcriptional control of fatty acid metabolism in *Escherichia coli*. *Journal of Biological Chemistry, 276*, 17373-17379.

SI Appendix

Protein cloning and expression.

Escherichia coli nanR (UniProt accession-P0A8W0) was commercially synthesized by GenScript and sub-cloned into expression vector pET28a. A truncated *nanR* fragment ($NanR^{33-263}$) was amplified using specific primers (forward: 5'-AAGGAGATATACATATGCTCTCCGAAATGGTGGAAGAAG-3'; reverse:5'-GTGCGGCCGCAAGCTTTTATTTCTTTTTGTTGGTGGTCTG-3') and cloned into pET28a using an In-Fusion HD Cloning Kit (Takara Bio USA) as per the manufacturer's instructions. The resulting recombinant plasmid was transformed into chemically-competent *E. coli* BL21(DE3) cells, which were then cultured in Luria-Bertani growth medium supplemented with kanamycin (30 µg mL^{-1}) at 37 °C with shaking at 220 rpm to an OD_{600} of ~0.6. Protein expression was induced by the addition of isopropyl β-D-1-thiogalactopyranoside to 1 mM and the temperature was lowered to 26 °C for incubation overnight.

Protein purification.

All purification steps for NanR and $NanR^{33-263}$ were conducted at 4 °C. Cells were harvested by centrifugation (Sorvall LYNX 4000 Superspeed) at 7,000 x *g* for 10 min. Cell pellets were resuspended in buffer A (20 mM Tris−HCl (pH 8.0), 150 mM NaCl), supplemented with Complete protease cocktail inhibitor (Roche), and lysed by sonication (Hielscher UP200S Ultrasonic Processor). Cell debris and insoluble material was pelleted by centrifugation at 32,000 x *g* for 30 min. As an initial purification step, protein was precipitated with 40% ammonium sulphate for 1 h at 4 °C. Protein was pelleted at 11,000 x *g* for 15 min and resuspended in buffer B (20 mM Tris−HCl (pH 8.0), 50 mM NaCl), while the supernatant was discarded. The resuspended sample was dialyzed overnight in buffer B. NanR was then purified using a three-step procedure: anion exchange, heparin affinity and size-exclusion chromatography using the ÄKTApure chromatography system (GE Healthcare). First, the dialyzed sample was applied to a HiTrap Q FF column (GE Healthcare) and washed with buffer B. Bound protein was eluted using a continuous gradient of buffer C (20 mM Tris−HCl (pH 8.0), 1 M NaCl). Fractions containing protein were identified by SDS-PAGE and subsequently pooled. The pooled sample was then applied to a HiTrap Heparin HP column (GE Healthcare), and the bound protein was eluted using a continuous gradient of buffer C. Following concentration to approximately 500 µL using Vivaspin 6 30 kDa molecular weight cutoff centrifugal concentrators (Sartorius), the pooled sample was applied to a Superdex 200 Increase 10/300 GL column (GE Healthcare) in buffer A, and NanR eluted as a single peak. The final purity was estimated to be approximately 95%, as highlighted by a single band in SDS-PAGE (Fig. 4*C*). Protein that was not immediately used in experiments was flash-frozen in liquid nitrogen and stored at −80 °C.

Double-stranded DNA formation.

Complementary DNA oligonucleotides (Integrated DNA Technologies) were resuspended in a buffer consisting of10 mM Tris−HCl (pH 8.0), and 100 mM NaCl, mixed at equimolar concentrations, and then hybridized by heating to 95 °C for 5 min, followed by cooling slowly to room temperature. DNA oligonucleotides used in EMSA and single-wavelength AUC experiments were 5'-FAM fluorescently-labelled on both strands to improve sensitivity. Double-stranded DNA oligonucleotides were stored at −20 °C until use.

Small-angle X-ray scattering (SAXS) analysis.

SAXS data were collected on the SAXS/WAXS beamline equipped with a Pilatus 1M detector (170 × 170 mm, effective pixel size 172 × 172 µm) at the Australian Synchrotron using a wavelength of 1.0332 Å. The sample detector distance was 1600 mm, providing a q range of 0.006−0.5 $Å^{-1}$. 80 µL aliquots of purified NanR and $NanR^{33-263}$ proteins at 10 mg mL^{-1} (340 µM) were injected onto an inline Superdex S200 5/150 Increase (GE Healthcare), equilibrated with buffer A, and supplemented with the radical scavenger 0.1% (w/v) sodium azide, using a flow rate of 0.45 mL min^{-1}. Protein-DNA complex was prepared by incubating NanR (340 µM) and two-binding-site DNA (170 µM) on ice for 30 min prior to injection. DNA alone was injected at the same concentration. To investigate the effect of Neu5Ac, the inline S200 column was re-equilibrated in buffer A and supplemented with 20 mM Neu5Ac. All scattering data was collected in 1-s exposures over a total of 500 frames using a 1.5-mm glass capillary at 12 °C. Coflow SAXS was used to minimize sample dilution and maximize signal-to-noise (1). 2D intensity plots were radially averaged, normalized against sample transmission, and background-subtracted using the Scatterbrain software package (Australian Synchrotron).

All scattering data was analyzed using the ATSAS software package (version 2.8.4) (2). PrimusQT (3) was used to perform the Guinier analysis. The pairwise distribution function $P(r)$ and the maximum interparticle dimension (D_{max}) of the scattering samples were calculated using an indirect Fourier transformation. The molecular mass of each sample was estimated using the SAXS-MoW2 package (4) and from the Porod volume. To generate a multiphase *ab initio* reconstruction of the dimeric NanR-DNA hetero-complex, 10 independent MONSA (5) runs were performed and then averaged using DAMAVER (6). Each model was further evaluated using SUPCOMB (7) to determine the level of consistency. The resulting averaged model was visualized in PyMOL (8) by increasing the solvent density radius. The scattering data for all samples, excluding DNA alone, were deposited into the Small Angle Scattering Biological Data Bank (https://www.sasbdb.org/) (*SI Appendix*, Table S1).

Single-wavelength sedimentation velocity experiments.

Sedimentation velocity experiments were performed using a Beckman Coulter Model XL-I analytical ultracentrifuge at University of Canterbury to analyze the oligomeric structure of NanR and the effect of Neu5Ac in solution. NanR at a concentration of 17 µM (0.5 mg mL^{-1}) in buffer A and in the presence of 20 mM Neu5Ac were loaded into a An-60 Ti four-hole rotor, using double sector epon-charcoal center-pieces, fitted with sapphire windows. Data were collected at 50,000 rpm and at 20 °C, where sedimentation was monitored using the absorbance optical system at 280 nm.

To assess protein-DNA interaction, sedimentation velocity experiments were performed in a Beckman Coulter Model XL-A analytical ultracentrifuge at University of Melbourne using double sector epon-charcoal center-pieces fitted with quartz windows in an An-50 Ti eight-hole rotor at 20 °C. NanR was titrated against a 5'-FAM fluorescently-labelled oligonucleotide at 11 protein concentrations (2-fold dilutions from 794–0.78 nM) in buffer A. The DNA oligonucleotide contained three repeats of the motif GGTATA (full recognition site). Data were collected at 50,000 rpm, where sedimentation was monitored using the fluorescence emission optical system (AVIV Biomedical). To generate an artificial bottom, 50 µL of FC43 fluorinert oil were loaded into the bottom of each cell. A radial calibration was performed prior to each experiment at 3,000 rpm using a calibration cell containing 10 µM fluorescein in 10 mM Tris–HCl (pH 7.8) and 100 mM KCl. Photo-multiplier tube (PMT) voltage and gain were adjusted for each cell, while an appropriate focusing depth was selected to maximize the signal and minimize the inner filter effect for the highest NanR concentration. A PMT voltage and gain setting of 58% was used.

All data were analyzed using SEDFIT (9). Sedimentation data was fitted to either a continuous size distribution [*c(s)*] or a continuous mass distribution [*c(M)*] model. The buffer density, buffer viscosity and an estimate of the partial specific volume of the protein sample based on the amino acid sequence were also determined using SEDNTERP.

Electrophoretic mobility shift assays.

Double stranded 5'-FAM-labelled DNA oligonucleotides were diluted to 10 nM in gel shift buffer (10 mM MOPS (pH 7.5), 50 mM KCl, 5 mM $MgCl_2$, and 10% (v/v) glycerol). Twelve-well Novex 6% Tris-glycine gels (Invitrogen) were pre-run in 0.5× Tris-Borate-EDTA (TBE) buffer (40 mM Tris-HCl (pH 8.3), 45 mM boric acid, and 1 mM EDTA) at 200 V for 30 min at 4 °C. Protein and DNA oligonucleotides were mixed and incubated at room temperature for 30 min to allow samples to reach equilibrium. Electrophoresis was performed immediately on the pre-run gels in 0.5× TBE buffer at 200 V for 20 mins at 4 °C. Following electrophoresis, gels were rinsed in distilled water and imaged using a Typhoon FLA 9500 Biomolecular Imager (GE Healthcare) with a 473 nm excitation source and a long-pass (LPB) emission filter.

Determining the dissociation constant of the NanR-DNA interaction.

The apparent affinity of the interaction between NanR and its DNA recognition sequence was measured by comparing the ratio of bound and unbound oligonucleotides in both the EMSA and analytical ultracentrifugation experiments. This ratio was determined from the fluorescent sedimentation velocity data by the integration of the peaks in the c(*s*) distribution, where a shifted species relative to the DNA signal represented complex formation. In the EMSA, the ratio of free versus bound fluorescently-labelled DNA was determined by densitometry using ImageJ (10). Together, the fraction containing bound DNA was plotted against the concentration of NanR in each experiment. The dissociation constant was calculated by fitting the data to the Hill equation using Prism 7 (GraphPad Software Inc.). The standard error of the best fit parameters was then calculated for the model (Fig. 2).

Multi-wavelength sedimentation velocity (MWL-SV) experiments.

To define the stoichiometry of the NanR-DNA complex, MWL-SV experiments were performed in a Beckman Coulter Optima AUC^{TM} using double sector epon-charcoal center-pieces fitted with sapphire windows in an An-60 Ti four-hole rotor at 20 °C. Samples were prepared with increasing loading concentrations of NanR with respect to DNA (1:1; 2:1; 3:1; 4:1; 6:1; 10:1 (protein:DNA)) in 50mM sodium phosphate (pH 7.4), and 50mM NaCl. Data were collected in intensity mode at either 50,000 or 60,000 rpm and sedimentation was monitored using the UV absorption system in intensity mode. This system allowed the acquisition of full UV-vis intensity spectra from 190–800 nm. In this experiment, radial sedimentation velocity scans were recorded in the range of 220–300 nm with 2 nm increments, providing 41 individual datasets for each loading concentration. All data were analyzed using UltraScan 4.0, release 2578 (11), and fit using the UltraScan LIMS cluster.

Initially all MWL-SV datasets were analysed using a two-dimensional spectrum analysis (2DSA) (12). This analysis method allows the sedimentation coefficient and frictional ratio (f/f_0) to be floated independently in an unbiased hydrodynamic model. In addition, both time-invariant and radially in-variant noise components are removed, while accurate fitting of the meniscus position is provided. All datasets were further refined using the 2DSA Monte Carlo (2DSA-MC) analysis, which determines the confidence intervals for each parameter that define each solute (13). These include the sedimentation and diffusion coefficients. The partial specific volume for NanR was predicted based on the amino acid sequence using UltraScan, while the buffer density and viscosity were determined based on the composition (50mM sodium phosphate (pH 7.4), 50mM NaCl) using UltraScan.

Absorbance spectra for NanR and DNA were obtained by performed a dilution series of each pure component across the spectral range of interest (220-300 nm) using a Genesys 10s benchtop spectrophotometer (Thermo Scientific). The dilution series of absorbance spectra were fitted to intrinsic

extinction profiles as described previously (14). The resulting intrinsic extinction profiles were scaled to molar concentration using an extinction coefficient of 13 980 M^{-1} cm^{-1} at 280 nm for full-length NanR and for $NanR^{32-263}$ as calculated by ExPASy ProtParam (15) from the amino acid sequence. For the DNA oligonucleotides, and extinction coefficient of 276 816 M^{-1} cm^{-1} at 260 nm for 2rDNA, and 567 112 M^{-1} cm^{-1} at 260 nm for 3rDNA as determined by the nearest-neighbour method (16; 17). The vector angle between these spectral profiles was found to be 60.2° and 63.3°, respectively for NanR and the 2rDNA or 3rDNA oligonucleotides. These vector angles represent good orthogonality between spectra and therefore ensure separability, as an angle of 0° reflects linear dependence or perfect overlap, while an angle of 90° indicates perfect orthogonality or no spectral overlap. Next, the profiles scaled to molar concentration were subsequently used to decompose the noise-corrected MWL-SV datasets into separate datasets for the protein and DNA components using the non-negatively constrained least squares algorithm (18) as previously described (14; 19). The resulting decomposed datasets were individually analysed by a two-dimensional spectrum analysis using UltraScan. The resulting amplitudes of the decomposed species involved in hetero-complex formation can be integrated to directly provide the molar stoichiometry of the protein-DNA hetero-complexes. A summary of the integration results from the MWL-SV analysis is shown in *SI Appendix* Table S3-5.

In order to accurately determine the molecular weight of each species in solution, a weight-averaged partial specific volume was estimated for each complex using the following equation:

$$\bar{\nu} = \frac{M_1\bar{\nu}_1 + M_2\bar{\nu}_2}{M_1 + M_2}$$

Here, the molar mass, measured in Daltons is required for the protein (M_1) and DNA (M_2) components, along with the partial specific volume of the protein ($\bar{\nu}_1$) and the DNA ($\bar{\nu}_2$). A molar mass of 59 kDa, 118 kDa and 177 kDa was used for the dimeric (oneNanR protomer), tetrameric (two NanR protomers) and hexameric (threeNanR protomers) protein components, respectively. A molar mass of 11.5 kDa and 21.5 kDa was used for the 2r- and 3r-DNA components, respectively. The partial specific volume of the protein ($\bar{\nu}_1$) was 0.7295 mL g^{-1}, while the partial specific volume of the DNA ($\bar{\nu}_2$) was 0.55 mL g^{-1}.

Crystallization, phase determination, and structure refinement.

Initial crystallisation conditions were identified using the sitting-drop vapour-diffusion method by mixing 400 nL of protein (20 mg mL^{-1} NanR in 20 mM Tris-HCl (pH 8.0), 150 mM containing 20 mM Neu5Ac) with 400 nL of reservoir solution equilibrated at 20 °C. However, following data collection at the Australian Synchrotron MX2 beamline, using an EIGER x 16M detector (Dectris), these crystals were not suitable for structure determination, diffracting to ~5 Å resolution. To overcome this, we performed *in situ* proteolysis (20) with the addition of 10 µg mL^{-1} chymotrypsin following identification of the

predominately disordered N-terminal extension (***SI Appendix*, Fig. S5**). We decided against surface-entropy reduction as this strategy utilises site-directed mutagenesis on charged residues (21). Since transcriptional regulators employ positive amino acids in DNA binding, we wanted to conserve these residues for modelling purposes.

Prior to crystallisation, the protein solution/protease mixture was incubated on ice for 30 min. After four weeks growth at 8 °C, crystals were produced, which diffracted to 2.10 Å using the sitting-drop vapor-diffusion method and *in situ* proteolysis by mixing 400 nL of NanR (20 mg mL^{-1}) NanR in buffer containing 20 mM Tris-HCl (pH 8.0), 150 mM , and 20 mM Neu5Ac) with 400 nL of reservoir solution (0.1 M sodium HEPES (pH 7.5), 0.2 M sodium acetate trihydrate, and 25% (w/v) PEG 3350), equilibrated at 8 °C. Crystals were flash-frozen in liquid nitrogen, and X-ray diffraction data were collected on the MX2 beamline at the Australian Synchrotron, using an EIGER × 16M detector (Dectris). A single dataset was collected at a wavelength of 0.9537 Å over 180° and was processed in XDS (22) displaying $P2_1$ symmetry, with cell dimensions of *a* = 39.77, *b* = 87.85, *c* = 73.87 Å. However, we were unable to phase the data using molecular replacement because the sequence identity of potential molecular replacement templates was low. To determine a phasing strategy, we performed X-ray absorption spectroscopy on the MX2 beamline to identify metal ions present in the crystal and obtained a distinct, yet weak, anomalous signal from zinc. The presence of an intrinsically-bound zinc ion was further supported using inductively-coupled plasma mass spectrometry. With the aim of increasing the anomalous signal, NanR was expressed in the presence of 100 µM $ZnCl_2$, purified, and subjected to further screening as described using *in situ* proteolysis. Crystals were obtained from a reservoir solution containing 0.1 M Tris-HCl (pH 8.5), 0.2 M magnesium chloride hexahydrate, and 30% (w/v) PEG 4000, flash-frozen, and then cryoprotected in the same reservoir solution supplemented with 15% (w/v) glycerol/ ethylene glycol. A MAD scan was performed on the crystal around the zinc absorption edge. At a wavelength of 1.2782 Å (slightly remote from the edge), 22 datasets were collected using an EIGER × 16M detector at a detector distance of 245–255 mm from a single crystal at five different positions where diffraction ranged from 2.62–2.29 Å. For each dataset, 3600 frames were collected with an exposure of 0.1 sec per frame, with an X-ray beam attenuation of 50%. These datasets were processed in XDS (22) displaying $I2_12_12_1$ symmetry and were then analyzed with XDS_NONISOMORPHISM (23) to identify the most isomorphous datasets. Based on this analysis, eight isomorphous datasets were selected, merged and scaled using XSCALE (22), in order to improve the zinc anomalous signal. The unit cell dimensions of the merged and scaled dataset were *a* = 75.46, *b* = 78.35, *c* = 88.72 Å, α=β=γ=90°. Data collection statistics are summarized in *SI Appendix*, Table S5.

The crystal structure of the C-terminal domain (residues 120–251) was solved using the SAD protocol in the Auto-Rickshaw pipeline (24).Input diffraction data were prepared and converted for use in Auto-Rickshaw using programs within the CCP4 suite (25). FA values were calculated using the program SHELXC (25). Based on an initial analysis of the data, the maximum resolution for substructure determination and initial phase calculation was set to 2.70 Å. Both heavy atoms requested were found using the program SHELXD (26). The correct hand for the substructure was determined using the ABS program (27), while initial phases of 162 were calculated following density modification using SHELXE (26). The initial phases were improved using density modification and phase extension to 2.29 Å resolution using RESOLVE (28). 50.00% of the model was built using the program ARP/wARP (29). This partial model, along with the 2.10 Å dataset, were used in a MRSAD protocol in Auto-Rickshaw (30) for phase enhancement and model completion. The resulting model was improved by iterative manual building in COOT (31) and refinement against the 2.10 Å dataset using PHENIX (32). No density was visible for residues 1–30 and 247–263 in chain A, or for residues 1–30 and 245–263 in chain B. The final refined model had R_{work} and R_{free} values of 18.2% and 23.0%, respectively. This model was validated using MOLPROBITY (33). Structure refinement statistics are summarized in Table *SI Appendix*, Table S5. The dimer interface was analyzed using PDBePISA (34). All structural graphics were prepared using PyMOL (8). The coordinates and structure factors for NanR in complex with Neu5Ac and zinc have been deposited in the PDB under the ID code 60N4.

Single-particle cryo-electron microscopy sample preparation, data acquisition and data processing.

Sample preparation:

Dimeric NanR-DNA hetero-complex was prepared by mixing NanR and the two-binding-site model DNA oligonucleotide a loading concentration of 2:1. Hexameric NanR-DNA hetero-complex was prepared by mixing NanR and the full recognition site DNA oligonucleotide at a loading concentration of 10:1. Following equilibration for 1 h at 4 °C, the sample was loaded onto a Superdex 200 Increase 10/300 GL column (GE Healthcare), pre-equilibrated with buffer A . Fractions consistent with complex formation were pooled and diluted to a final concentration of 0.5 mg mL^{-1}. Frozen-hydrated samples were prepared on plasma-cleaned Quantifoil R1.2/1.3 holey carbon EM grids (Quantifoil) using a Vitrobot Mark III vitrification robot (FEI) with a 3 sec blotting time, 100% humidity and -3 mm blotting offset.

Data collection:

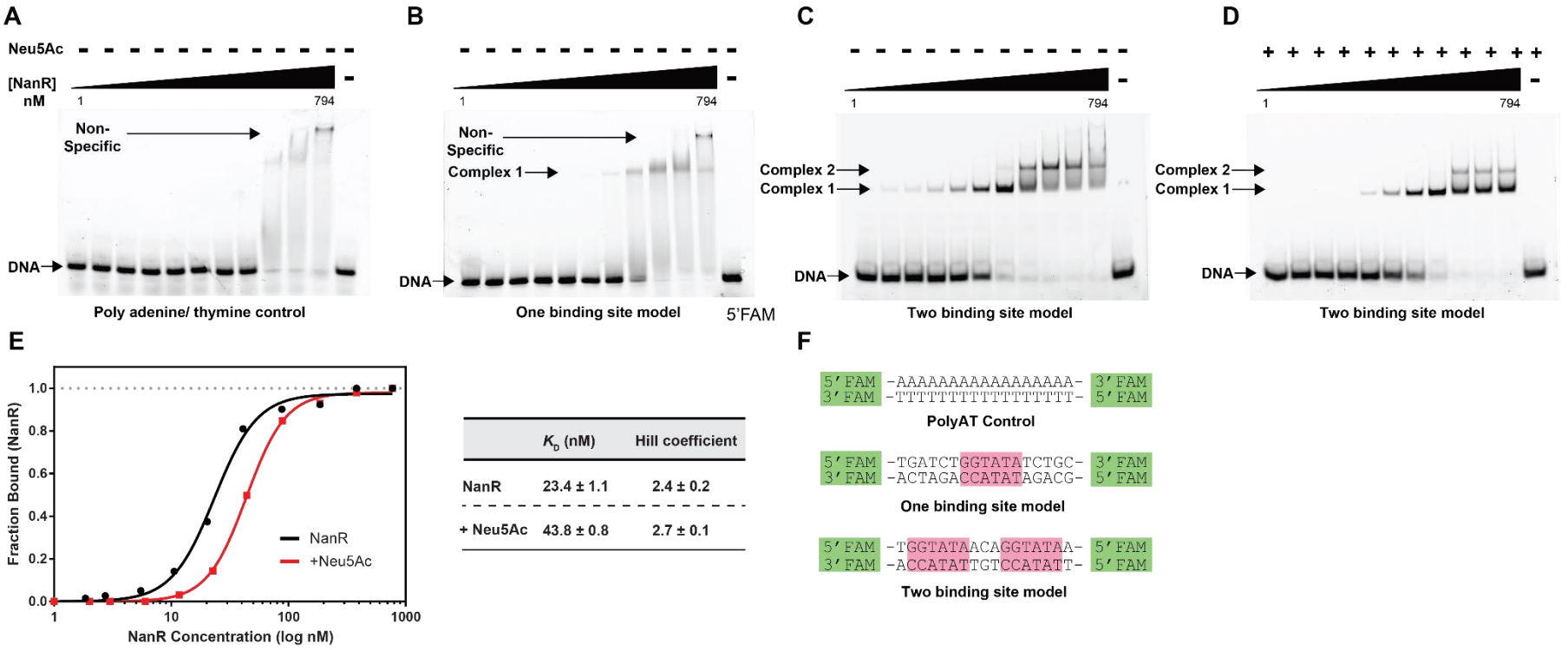
Automated data acquisition of the dimeric NanR-DNA hetero-complex was performed using a Titan Krios™ electron microscope (FEI) at 300 keV, which was equipped with a K2 Summit™ direct detector (Gatan) and a GIF- Quantum energy filter (Gatan) using EPU software (ThermoFisher Scientific). Cryo-

EM imaging were performed using nanoprobe EFTEM imaging mode, a C2 Condenser aperture size of 50 µM, an objective aperture size of 70 µM, and at a nominal magnification of 215,000×, which provided a magnified pixel size of 0.68 Å. Movies comprising 32 frames, an exposure time of 12.8 sec, and a total dose of 60e^{-} $Å^{-2}$ were collected using a K2 Summit™ direct detector (Gatan) in counting mode, using a dose rate of 4e^{-} $pixel^{-1}$ sec^{-1} and a defocus range from -0.5 µm to -2.5 µm. Automated data acquisition of the hexameric NanR-DNA hetero-complex was performed using a Talos Artica™ electron microscope (FEI) at 200 keV, which was equipped with a Falcon III ™ direct detector (FEI).

Data processing:

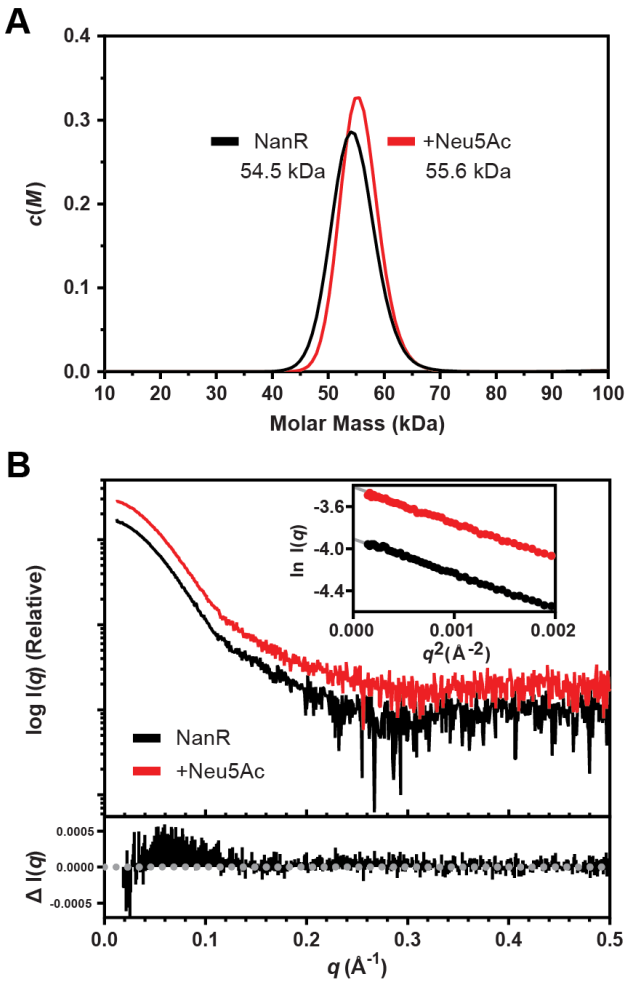
Correction for beam-induced motion were performed using MotionCor2, which provided dose weighted corrected averages. CTF parameters were estimated on the corresponding non-dose weighted averages using Gctf v1.06 (35). Both steps were performed using RELION v3.0-stable wrapper. Automated particle-picking were performed using Gautomatch v0.53 on a subset of images with a sphere diameter of 8 nm. These particles were then subjected to 2D classification, where the 2D classes that showed secondary structural features were used as a template for further automated particle-picking using the entire dataset. This generated 1.2m particles. Further 2D classification resulted in 270k particles with good signal-to-noise. These particles were then used to generate the initial 3D model using *ab initio* reconstruction in CRYOSPARC v2 (36). Further 3D classifications in RELION 3.0 teased out a homogenous set of 140k particles corresponding to the DNA bound dimer population. Multiple rounds of masked refinement resulted in teasing out the signal of 70kDa NanR-DNA hetero-complex at a resolution of 3.9 Å (gold standard FSC 0.143 criteria). The 3D structure of the DNA oligonucleotide was modelled using the 3D-DART server (37). The model was fit into the map using *Cryo_fit* within PHENIX (38).

SI Appendix Fig S1 | Electrophoretic mobility shift assay (EMSA) of NanR and fluorescently-labelled DNA oligonucleotides. *(**A**)* Specific DNA binding is abrogated in the poly adenine/ thymine (poly-AT) control. At high concentration, non-specific binding is observed. *(**B**)* Poor DNA binding activity is observed when only one GGTATA repeat is present suggesting the regulation mechanism is dependent on cooperative binding. *(**C**)* Two concentration-dependent complexes are observed when two GGTATA repeats are present. *(**D**)* A small change in DNA binding affinity is observed in the presence of 20 mM Neu5Ac when compared with the EMSA gel in the absence of Neu5Ac *(C)*. *(**E**)* The binding isotherm from the two-binding-site model with NanR (black) and in the presence of Neu5Ac (red). The data was best fit to a Hill equation due to the signal dependency. A small decrease in affinity is observed (~20 nM). *(**F**)* The sequence of the DNA oligonucleotides used in EMSA experiments. The fluorescent fluorescein (FAM) tag (green) and each GGTATA repeat (pink) is highlighted.

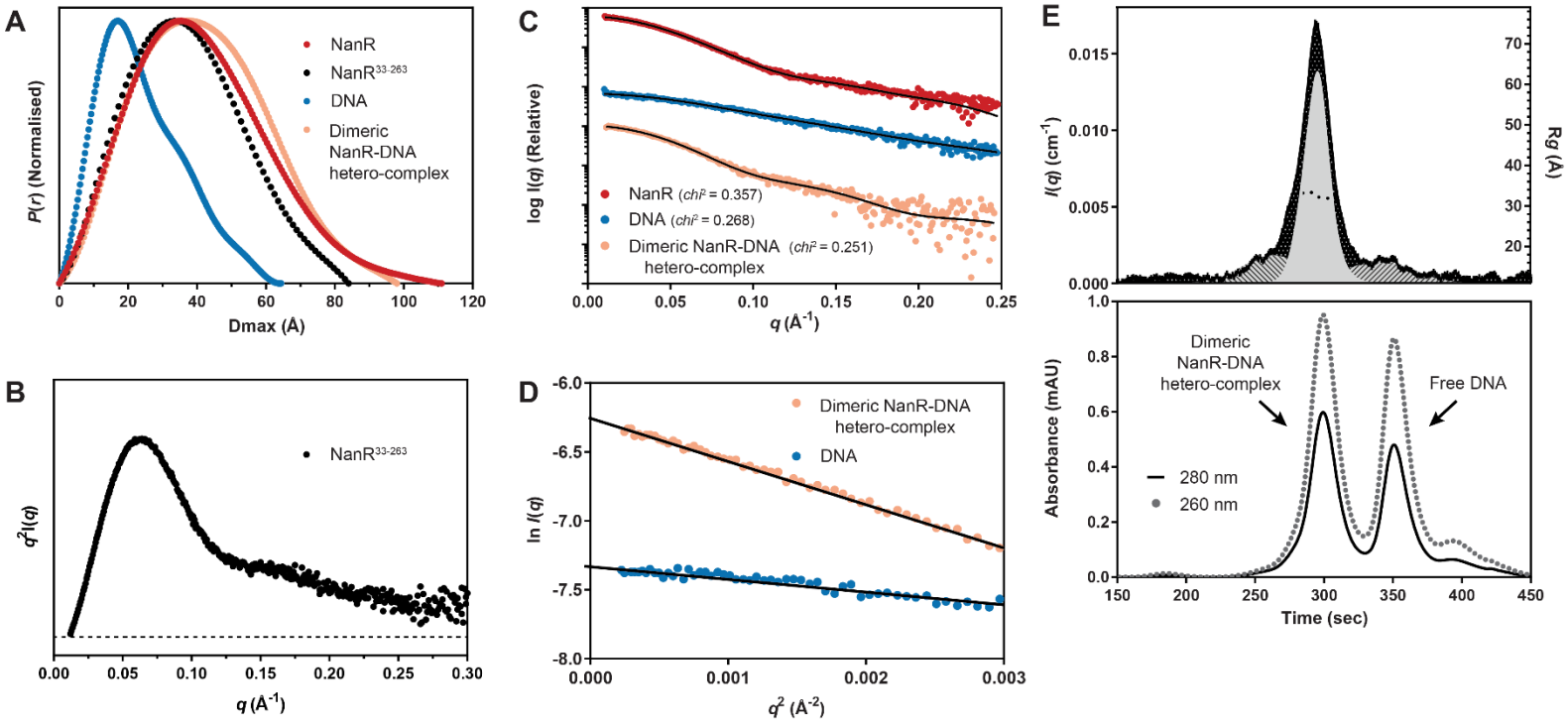


	K_D (nM)	Hill coefficient
NanR	23.4 ± 1.1	2.4 ± 0.2
+ Neu5Ac	43.8 ± 0.8	2.7 ± 0.1

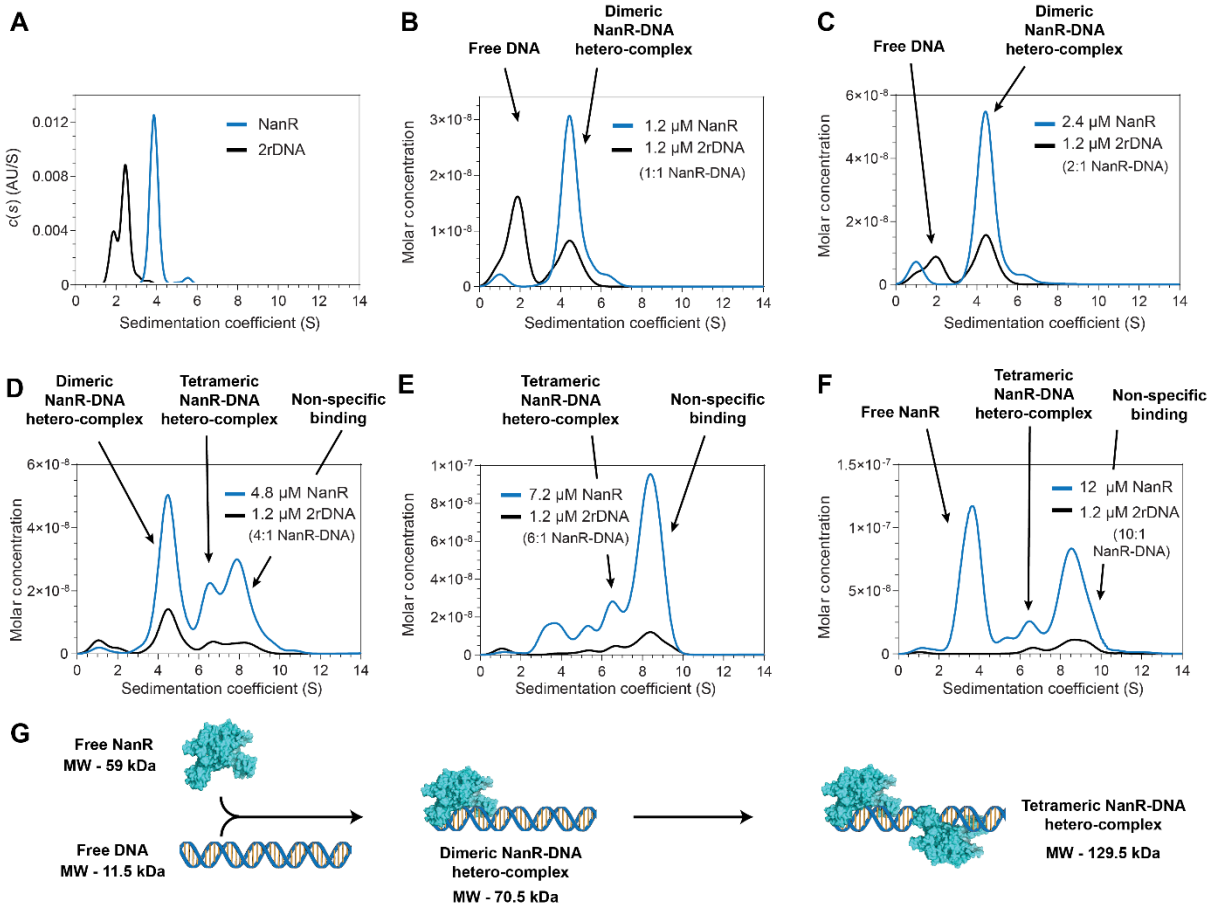
SI Appendix* Fig. S2 |** *E. coli* NanR remains as an exclusive dimer in solution with (red) or without (black) the presence of Neu5Ac. (A***) Sedimentation velocity data at a protein concentration of 17 μM and with the addition of 20 mM Neu5Ac were fitted to a continuous mass [c(*M*)] distribution. The data were both best fit to a dimeric species with a calculated mass of 59 kDa. (***B***) Small angle X-ray scattering data present an analogous scattering profile with and without Neu5Ac (20 mM), demonstrating no oligomeric change has occurred in solution. The residuals (shown below) highlights the *q* region where a minor difference is observed—this may be reflective of a small Neu5Ac-induced conformational change. The linearity of the Guinier plot (insert) further supports the observation of a single monodisperse species in solution, free of any significant amount of aggregation or inter-particle interference.



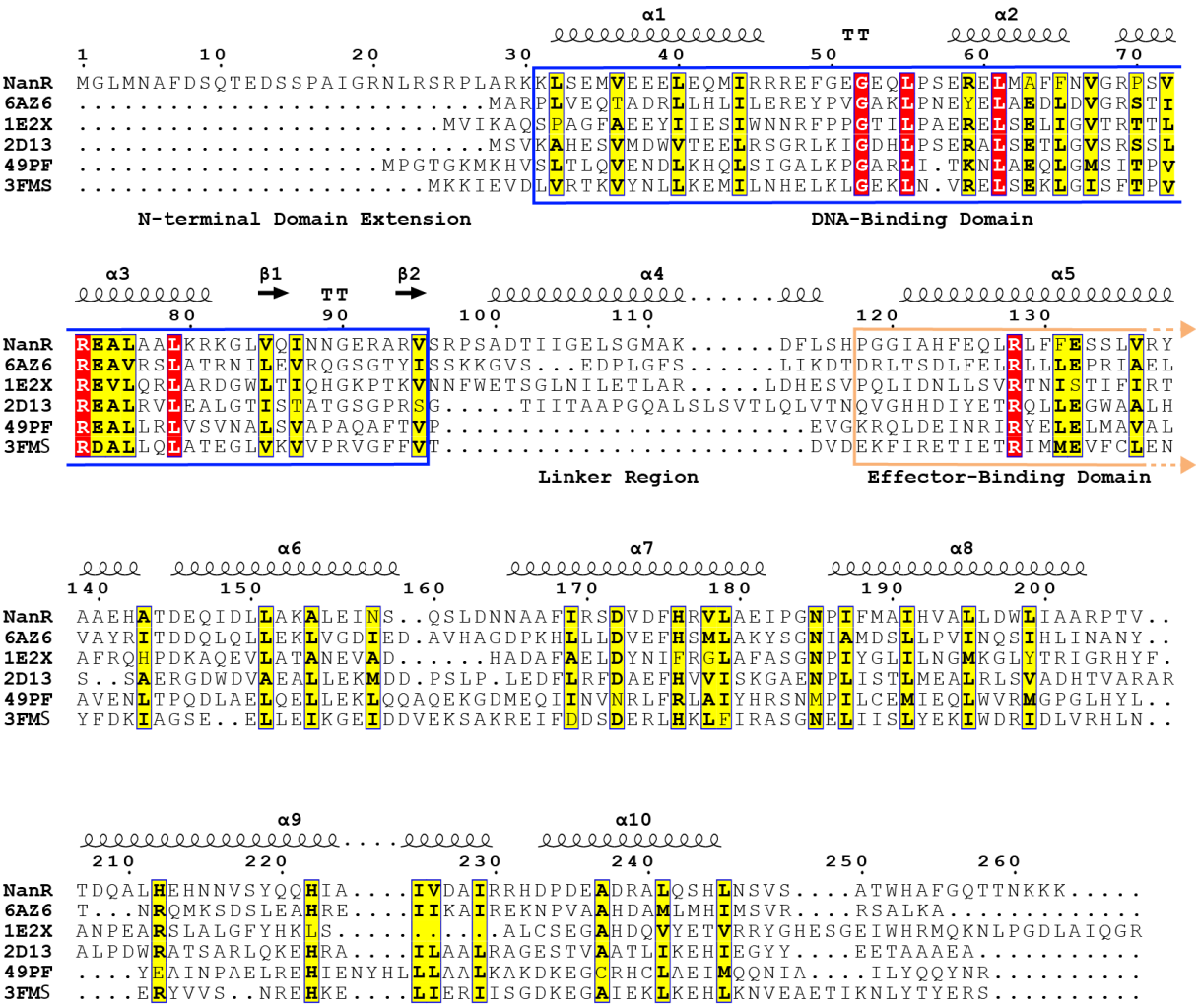
SI Appendix* Fig. S3 |** SAXS for *E. coli* NanR and the NanR-DNA hetero-complex. (A***) Pairwise distribution plot displaying the maximum interparticle dimension (D_{max}) for NanR (111 Å, red), $NanR^{33-263}$ (84 Å, red), two binding site DNA (64 Å, blue) and the dimeric NanR-DNA hetero-complex (98 Å, orange). (***B***) Kratky plot for $NanR^{33-263}$ presents a bell-shaped curve, which indicates the protein is folded in solution. In addition, the Kratky analysis provided evidence of flexibility in the structure as the plot does not converge with the baseline (black dash). (***C***) The scattering profiles for NanR (red), DNA (blue) and the dimeric NanR-DNA hetero-complex (orange). The MONSA generated fits (black line) and chi^2 value are shown for each. (***D***) The Guinier plot for DNA (blue) and the dimeric NanR-DNA hetero-complex (orange) highlights linearity at low q—this provides confidence that the data is free of any significant amount of aggregation or interparticle interference. (***E***) The SEC-SAXS data for the dimeric NanR-DNA hetero-complex. Using US-SOMO (39), three Gaussian functions were fitted to the data to deconvolute the scattering profile for the dimeric NanR-DNA hetero-complex (peak 2, grey) from free DNA and aggregate (peak 1 and 3, diagonal line). The radius of gyration (R_g) value across peak 2 is shown as black dots. The lower panel presents the UV trace for the SEC-SAXS data collected at 260 nm (grey dot), and 280 nm (black line). The NanR-DNA hetero-complex and free DNA are highlighted.



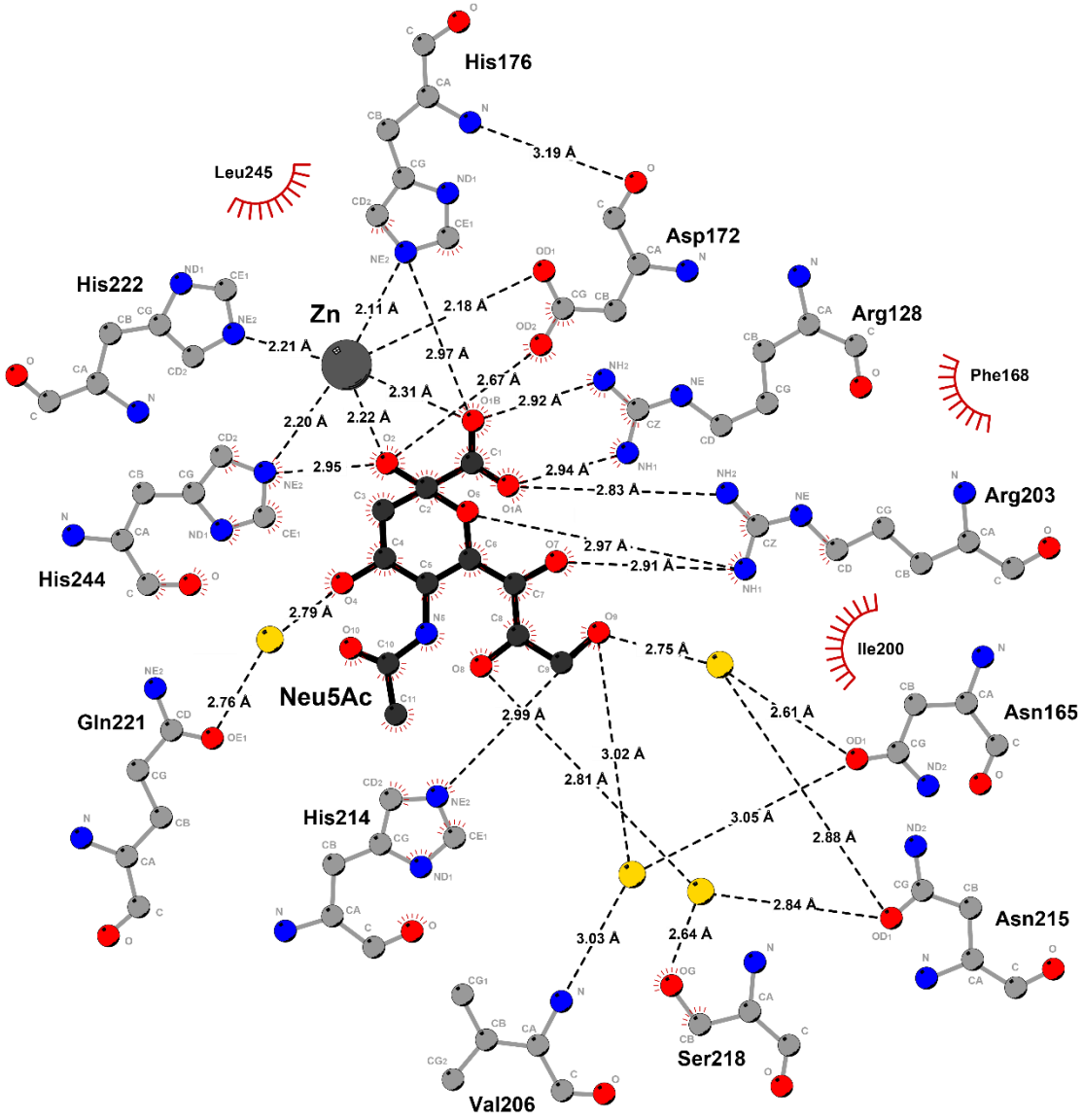
SI Appendix Fig S4 | The stoichiometry of the NanR-DNA hetero-complex**.** The deconvoluted sedimentation coefficient distributions for the titration series are presented. (***A***) Separate controls of NanR (blue) and the two-binding-site DNA model (2rDNA, black). (***B***) 1.2 µM NanR. (***C***) 2.4 µM NanR. (***D***) 4.8 µM NanR. (***E***) 7.2 µM NanR. (***F***) 12 µM NanR$^{33\text{-}263}$. Upon increasing protein concentration, a clear shift in the sedimentation coefficient is observed, consistent with the formation of higher order hetero-complexes. The broad reaction boundary is indicative of tight binding and represents multiple, dynamic intermediate species that are mediated by mass action. The presence of excess protein, free of any co-migrating DNA indicates that hetero-complex formation has reached saturation. All plots are presented as *g*(*s*) distributions. (***G***) A model of the hetero-complex assembly process is presented.



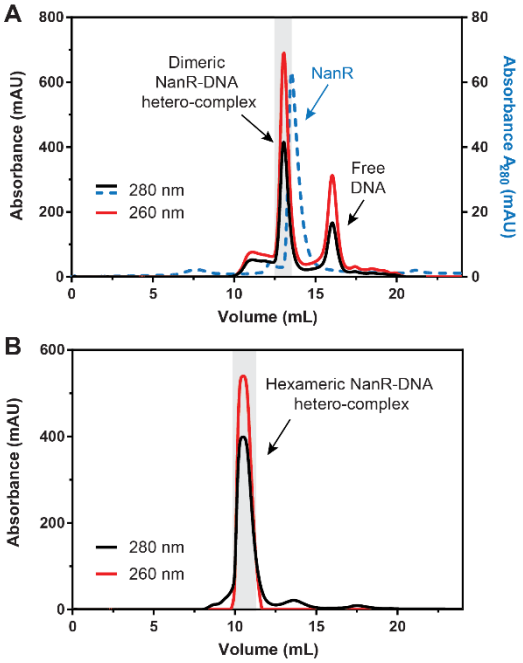
SI Appendix Fig S5 | Multiple sequence alignment performed using Clustal Omega (40) and generated using ESPript 3.0 (41). The top five hits from a sequence homology search within the PDB were used in the sequence alignment. The red background highlights identical residues, while the yellow background indicates chemically similar residues. Secondary structure elements depicted above the alignment were generated from the crystal structure of *E. coli* NanR (Fig. 4*A*). The dots above the aligned sequences are the amino acid numbers respective to NanR. The N-terminal DNA-binding domain (blue box), the linker region, and the C-terminal effector-binding domain (beige box).



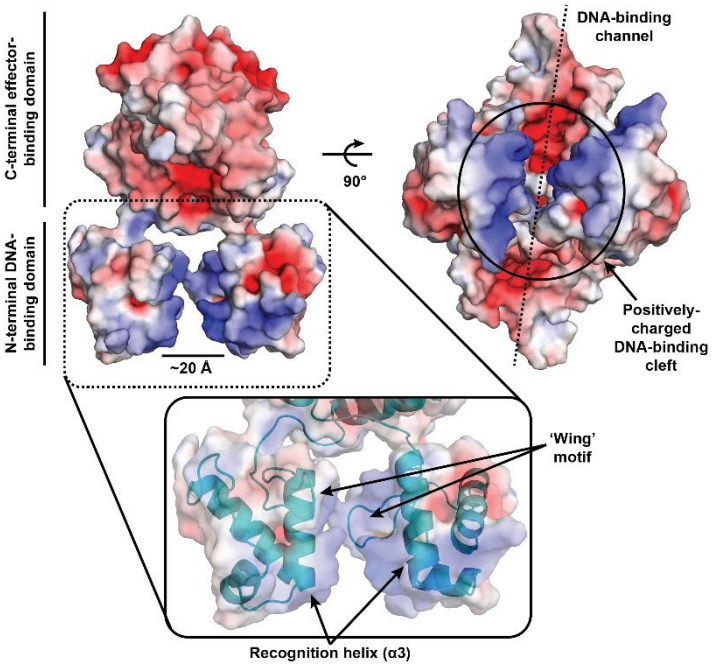
SI Appendix Fig S6 | Schematic representation of the NanR-Neu5Ac-Zn interaction network. The hydrogen bonded network is presented (black dashed lines) and the distance between partners are labelled. Key interacting water molecules (yellow spheres) and hydrophobic contacts (red arcs with spokes) are shown. Figure generated by LigPlot+ (42).



SI Appendix Fig S7 | Cryo-EM sample preparation. Prior to vitrification, dimeric (***A***) and hexameric (***B***) NanR-DNA hetero-complex was purified using size exclusion chromatography to remove free DNA and aggregates. The pooled fraction is highlighted (grey box). The size exclusion chromatogram for NanR (blue dashed lines) in the absence of DNA is presented in (*A*).



SI Appendix Fig S8 | The electrostatic surface map for NanR. The electrostatic surface map was generated using APBS (43) and coloured by the electrostatic potential distribution with negatively and positively charged regions displayed in red and blue, respectively (range -5kT to 5kT). Located between the N-terminal DNA-binding domains is a positively-charged cleft that is proposed to be the DNA binding site. The distance between the N-terminal domains is ~20 Å. The most positive electrostatic surface (shown as an insert) is mapped to recognition helix (α3) and the 'wing' of the winged helix-turn-helix motif.



SI Appendix Table S1 | Data collection and analysis statistics from solution studies.

Sedimentation velocity analysis	NanR	NanR (+ 20 mM Neu5Ac)			
Sedimentation coefficient (×10^{-13} S)	3.65	3.55			
Frictional ratio (f/f_0)	1.34	1.31			
Molecular weight (from c(M), kDa)	54.5	55.6			
Partial specific volume of NanR (mL g^{-1})	0.7295				
Buffer density (g/cm^3)	1.006	1.009			
Buffer viscosity (cp)	1.027	1.086			
SAXS data analysis	**NanR**	**NanR (+ 20 mM Neu5Ac)**	**$NanR^{33-263}$**	**2rDNA**	**Dimeric NanR-DNA hetero-complex**
$I(0)$ (cm^{-1}) (from Guinier analysis)	0.021 ± 5.7e-05	0.032 ± 8.7e-03	0.057 ± 6.2e-05	0.006 ± 4.2e-05	0.013 ± 6.7e-05
R_g (Å) (from Guinier analysis)	32.3 ± 0.3	31.4 ± 0.2	28.9 ± 0.1	19.7 ± 0.5	33.4 ± 0.6
$I(0)$ (cm^{-1}) (from $P(r)$ analysis)	0.021 ± 0.1e-03	0.033 ± 0.1e-03	0.057 ± 0.9e-04	0.006 ± 0.8e-04	0.013 ± 0.1e-03
R_g (Å) (from $P(r)$ analysis)	31.8 ± 0.2	32.3 ± 0.2	28.50 ± 0.4	19.8 ± 0.3	32.9 ± 0.3
D_{max} (Å)	111	116	84	64	98
Porod volume ($Å^{-3}$)	112 000	107 000	83 000	16 000	114 000
Molar mass (from Porod Volume, kDa)	65.9	62.9	48.8	9.4	67.1
Molar mass (from SAXSMoW2*, kDa)	70.6	63.6	57.7	10.6	82.5
Calculated dimer MM from sequence (kDa)	59.0	59.0	51.3	11.5	70.5
SASBDB ID	SASDFJ5	SASDFK5		-	
SAXS data-collection parameters					
Instrument	Australian Synchrotron SAXS/WAXS beamline				
Detector	PILATUS 1M (Dectris)				
Wavelength (Å)	1.0332				
Maximum flux at sample	8 x 10^{12} photons per second at 12 keV				
Camera length (mm)	1600				
q range ($Å^{-1}$)	0.006-0.5				
Exposure time	Continuous 1 second frame measurements				
Sample configuration	SEC-SAXS with co-flow				
Sample temperature (°C)	12				

[a] http://saxs.ifsc.usp.br/ (4)

SI Appendix Table S2 I Integration results from MWL-SV analysis of separated NanR and two-binding-site DNA (2rDNA) signals

	NanR	2rDNA	NanR:2rDNA						
			1:1	2:1	4:1 (Species 1)	4:1 (Species 2)	6:1 (Species 1)	6:1 (Species 2)	10:1
Sed. Coefficient (x10^{-13} S) *	3.76 (3.51, 4.01)	2.01 (1.85, 2.18)	4.39 (4.13, 4.57)	4.41 (3.91, 4.91)	4.40 (4.26, 4.55)	6.51 (5.87, 7.14)	6.26 (5.99, 6.54)	8.61 (8.34, 8.87)	8.78 (8.00, 9.56)
Dif. Coefficient (x10^{-07} D) *	6.3 (3.84, 8.76)	7.26 (5.41, 9.12)	4.13 (3.67, 4.59)	5.1 (4.30, 5.89)	4.49 (3.51, 5.47)	N/D†	N/D†	3.74 (2.19, 5.3)	4.24 (4.16, 4.32)
Measured molar ratio ‡	n/a	n/a	2.10	2.84	2.58	4.02	4.23	6.33	6.48
Weight-averaged $\bar{\nu}$ (mL g^{-1}) §	0.730	0.55	0.700	0.700	0.700	0.714	0.714	0.719	0.719
Measured molar mass (kDa) ¶	54.0 (30.0, 77.9)	14.9 (9.98, 19.98)	86	73.1	79.4	N/D†	N/D†	148.6	178.6
Theoretical molar mass (kDa) #	59	11.5	70.5	70.5	70.5	129.5	129.5	188.5	188.6
Oligomeric state of hetero-complex	n/a	n/a	Dimeric	Dimeric	Dimeric	Tetrameric	Tetrameric	Hexameric	Hexameric

* These are the sedimentation and diffusion coefficients observed following 2DSA-Monte Carlo analysis. Parameters are obtained from integration of pseudo-3D plots within the UltraScan software. All measured values represent the mean from the Monte Carlo analysis. The values in parentheses are the 95% confidence intervals from the Monte Carlo analysis.

† N/D = not determined as species were present within a reaction boundary.

‡ Partial concentration is determined from peak integration of the co-migrating species in both the NanR and DNA datasets. As the data is scaled to molar concentration these concentrations can be used directly to infer the molar ratio and thus stoichiometry of the complex.

§ Partial specific volume ($\bar{\nu}$) of the complex, estimated from the weight-average of the protein and DNA components. A $\bar{\nu}$ of 0.7295 mL g^{-1} was used for NanR, while a $\bar{\nu}$ of 0.55 mL g^{-1} was used for DNA. The equation used to calculate the weight-averaged $\bar{\nu}$ is presented in *SI Methods*.

¶ The measured molar mass is estimated based upon the hydrodynamic parameters (sedimentation and diffusion coefficient) and the weighted-averaged $\bar{\nu}$, predicted based on the measured molar ratio. The values in parentheses are the 95% confidence intervals from the Monte Carlo analysis for the pure components.

The theoretical mass is predicted based upon the amino acid or nucleic acid sequence within UltraScan. The mass of each complex is predicted based on the observed molar ratio.

SI Appendix Table S3 | Integration results from MWL-SV analysis of separated NanR and full operator site DNA (3rDNA) signals

			NanR:3rDNA						
	NanR	3rDNA	1:1 (Species 1)	1:1 (Species 2)	3:1 (Species 1)	3:1 (Species 2)	6:1 (Species 1)	6:1 (Species 2)	10:1
Sed. Coefficient (x10^{-13} S) *	3.72 (3.50, 3.94)	2.76 (2.63, 2.89)	5.67 (5.09 6.26)	7.45 (7.01, 7.90)	5.62 (4.83, 6.41)	7.46 (6.82, 8.11)	7.54 (7.13, 7.95)	7.99 (7.35, 8.61)	8.31 (7.95, 8.68)
Dif. Coefficient (x10^{-07} D) *	6.3 (3.84, 8.76)	6.99 (6.10, 7.86)	4.37 (2.46, 6.30)	N/D†	4.40 (2.18, 6.11)	N/D†	N/D†	3.16 (2.65, 3.66)	3.40 (2.49, 4.31)
Measured molar ratio ‡	n/a	n/a	2.45	4.25	2.37	3.63	4.65	6.44	6.21
Weight-averaged $\bar{\nu}$ (mL g^{-1}) §	0.730	0.550	0.682	0.702	0.682	0.702	0.702	0.710	0.710
Measured molar mass (kDa) ¶	54.0 (30.0, 77.9)	21.4 (17.9, 24.8)	98.9	N/D†	97.6	N/D†	N/D†	211.8	204.8
Theoretical molar mass (kDa) #	59.0	21.5	80.5	139.5	80.5	139.5	139.5	198.5	198.5
Oligomeric state of hetero-complex	n/a	n/a	Dimeric	Tetrameric	Dimeric	Tetrameric	Tetrameric	Hexameric	Hexameric

* These are the sedimentation and diffusion coefficients observed following 2DSA-Monte Carlo analysis. Parameters are obtained from integration of pseudo-3D plots within the UltraScan software. All measured values represent the mean from the Monte Carlo analysis. The values in parentheses are the 95% confidence intervals from the Monte Carlo analysis.

† N/D = not determined as species were present within a reaction boundary.

‡ Partial concentration is determined from peak integration of the co-migrating species in both the NanR and DNA datasets. As the data is scaled to molar concentration these concentrations can be used directly to infer the molar ratio and thus stoichiometry of the complex.

§ Partial specific volume ($\bar{\nu}$) of the complex, estimated from the weight-average of the protein and DNA components. A $\bar{\nu}$ of 0.7295 mL g^{-1} was used for NanR, while a $\bar{\nu}$ of 0.55 mL g^{-1} was used for DNA. The equation used to calculate the weight-averaged $\bar{\nu}$ is presented in *SI Methods*.

¶ The measured molar mass is estimated based upon the hydrodynamic parameters (sedimentation and diffusion coefficient) and the weighted-averaged $\bar{\nu}$, predicted based on the measured molar ratio. The values in parentheses are the 95% confidence intervals from the Monte Carlo analysis for the pure components.

The theoretical mass is predicted based upon the amino acid or nucleic acid sequence within UltraScan. The mass of each complex is predicted based on the observed molar ratio.

SI Appendix Table S4 | Integration results from MWL-SV analysis of separated NanR$^{33-263}$ and two-binding-site DNA (2rDNA) signals

	NanR (33-263)	3rDNA	NanR:3rDNA	
			3:1	10:1
Sed. Coefficient (x10^{-13} S) *	3.42 (3.09, 3.77)	2.76 (2.63, 2.89)	4.39 (4.10, 4.67)	4.46 (3.58, 5.33)
Dif. Coefficient (x10^{-07} D) *	6.79 (4.09, 9.49)	6.99 (6.10, 7.86)	4.97 (4.03, 5.91)	5.04 (3.39, 6.70)
Measured molar ratio ‡	n/a	n/a	2.44	2.24
Weight-averaged $\bar{\nu}$ (mL g^{-1}) §	0.730	0.55	0.677	0.677
Measured molar mass (kDa) ¶	45.3 (29.7, 62.8)	21.4 (17.9, 24.8)	66.3	66.4
Theoretical molar mass (kDa) #	51.3	21.5	72.8	72.8
Oligomeric state of hetero-complex	n/a	n/a	Dimeric	Dimeric

* These are the sedimentation and diffusion coefficients observed following 2DSA-Monte Carlo analysis. Parameters are obtained from integration of pseudo-3D plots within the UltraScan software. All measured values represent the mean from the Monte Carlo analysis. The values in parentheses are the 95% confidence intervals from the Monte Carlo analysis.

‡ Partial concentration is determined from peak integration of the co-migrating species in both the NanR and DNA datasets. As the data is scaled to molar concentration these concentrations can be used directly to infer the molar ratio and thus stoichiometry of the complex.

§ Partial specific volume ($\bar{\nu}$) of the complex, estimated from the weight-average of the protein and DNA components. A $\bar{\nu}$ of 0.7295 mL g^{-1} was used for NanR, while a $\bar{\nu}$ of 0.55 mL g^{-1} was used for DNA. The equation used to calculate the weight-averaged $\bar{\nu}$ is presented in *SI Methods*.

¶ The measured molar mass is estimated based upon the hydrodynamic parameters (sedimentation and diffusion coefficient) and the weighted-averaged $\bar{\nu}$, predicted based on the measured molar ratio. The values in parentheses are the 95% confidence intervals from the Monte Carlo analysis for the pure components.

The theoretical mass is predicted based upon the amino acid or nucleic acid sequence within UltraScan. The mass of each complex is predicted based on the observed molar ratio.

SI Appendix Table S5 | Data collection and refinement statistics from X-ray crystallography*

Dataset name	NanR-Zn	NanR
Data collection		
X-ray wavelength (Å)	1.2781	0.9537
Crystal-to-detector distance (mm)	245	200
Space group	$I\ 2_1\ 2_1\ 2_1$	$P\ 2_1$
Unit cell *a*, *b*, *c* (Å)	75.46/78.35/88.72	39.77, 87.85, 73.87
Unit cell α, β, γ (°)	90, 90, 90	90, 103, 90
Resolution range (Å)	44.36-2.29	43.9-2.10 (2.17-2.10)
No. of total reflections	12209	96673
No. of unique reflections	7465	28482
Completeness (%)	99.6 (97.9)	98.1 (88.3)
R_{merge} (%)	5.4 (36.9)	4.4 (53.5)
$CC_{1/2}$ (%)	99.9 (94.9)	99.9 (69.4)
$I/\sigma\ (I)$	4.23	15.3 (1.6)
Refinement statistics		
Resolution (Å)		43.9-2.10
No. of reflections used		26874
No. of reflection used in test set		1368
R_{work}/R_{free} (%)		18.12/22.95
Total non-H atoms		3653
Protein/ Water		3586
Ligands/ Ion		67
Mean *B* factor (Å^2)		56.65
Geometric RMSD		
Bond (Å)		0.009
Angle (°)		1.27
MolProbity statistics†		
Rotamer outliers (%)		3.88
Clashscore		9.17
Ramachandran plot		
Favoured/allowed regions (%)		98.36/1.64
PDB ID		60N4

* Values in parentheses are for the highest resolution shell

† *MolProbity* is a structure-validation web service (33)

SI Appendix References

1. Kirby, N., Cowieson, N., Hawley, A. M., Mudie, S. T., McGillivray, D. J., Kusel, M., Samardzic-Boban, V., & Ryan, T. M. (2016). Improved radiation dose efficiency in solution SAXS using a sheath flow sample environment. *Acta Crystallographica Section D-Structural Biology, 72*, 1254-1266.
2. Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., Kikhney, A. G., Hajizadeh, N. R., Franklin, J. M., Jeffries, C. M., & Svergun, D. I. (2017). ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography, 50*, 1212-1225.
3. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., & Svergun, D. I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography, 36*, 1277-1282.
4. Fischer, H., Neto, M. D., Napolitano, H. B., Polikarpov, I., & Craievich, A. F. (2010). Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *Journal of Applied Crystallography, 43*, 101-109.
5. Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biochemical Journal, 76*, 2879-2886.
6. Volkov, V. V., & Svergun, D. I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. *Journal of Applied Crystallography, 36*, 860-864.
7. Kozin, M. B., & Svergun, D. I. (2001). Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography, 34*, 33-41.
8. DeLano, W. L. (2002). The PyMOL Molecular Graphics Program. San Carlos: Delano Scientific.
9. Schuck, P. (2000). Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophysical Journal, 78*, 1606-1619.
10. Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods, 9*, 671-675.
11. Demeler, B. (2005). UltraScan - A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments. In D. J. Scott, S. E. Harding, & A. J. Rowe (Eds.), *Analytical Ultracentrifugation: Techniques and Methods* (pp. 210-230): The Royal Society of Chemistry.
12. Brookes, E., Cao, W. M., & Demeler, B. (2010). A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *European Biophysics Journal with Biophysics Letters, 39*, 405-414.
13. Demeler, B., & Brookes, E. (2008). Monte Carlo analysis of sedimentation experiments. *Colloid and Polymer Science, 286*, 129-137.
14. Gorbet, G. E., Pearson, J. Z., Demeler, A. K., Colfen, H., & Demeler, B. (2015). Next-Generation AUC: Analysis of Multiwavelength Analytical Ultracentrifugation Data. In J. L. Cole (Ed.), *Analytical Ultracentrifugation* (Vol. 562, pp. 27-47).
15. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research, 31*, 3784-3788.
16. Fasman, G. (1975). In Handbook of Biochemistry and Molecular Biology, Vol. I, Nucleic Acids, Chem. *Rubber Co., Cleveland, OH*, 76-206.
17. Cantor, C. R., Warshaw, M. M., & Shapiro, H. (1970). Oligonucleotide interactions. 3. Circular dichroism studies of the conformation of deoxyoligonucleotides. *Biopolymers, 9*, 1059-1077.
18. Lawson, C. L., & Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs, NJ., Prentice-Hall, Inc.
19. Zhang, J., Pearson, J. Z., Gorbet, G. E., Colfen, H., Germann, M. W., Brinton, M. A., & Demeler, B. (2017). Spectral and Hydrodynamic Analysis of West Nile Virus RNA-Protein Interactions by Multiwavelength Sedimentation Velocity in the Analytical Ultracentrifuge. *Analytical Chemistry, 89*, 862-870.
20. Dong, A., Xu, X., Edwards, A. M., Chang, C., Chruszcz, M., Cuff, M., Cymborowski, M., Di Leo, R., Egorova, O., Evdokimova, E., Filippova, E., Gu, J., Guthrie, J., Ignatchenko, A., Joachimiak, A., Klostermann, N., Kim, Y., Korniyenko, Y., Minor, W., Que, Q., Savchenko, A., Skarina, T., Tan, K., Yakunin, A., Yee, A., Yim, V., Zhang, R., Zheng, H., Akutsu, M., Arrowsmith, C., Avvakumov, G. V., Bochkarev, A., Dahlgren, L. G., Dhe-Paganon, S., Dimov, S., Dombrovski, L., Finerty, P., Jr., Flodin, S., Flores, A., Graslund, S., Hammerstrom, M., Herman, M. D., Hong, B. S., Hui, R., Johansson, I., Liu, Y., Nilsson, M., Nedyalkova, L., Nordlund, P., Nyman, T., Min, J., Ouyang, H., Park, H. W., Qi, C., Rabeh, W., Shen, L., Shen, Y., Sukumard, D., Tempel, W., Tong, Y., Tresagues, L., Vedadi, M., Walker, J. R., Weigelt, J., Welin, M., Wu, H., Xiao, T., Zeng, H., & Zhu, H. (2007). In situ proteolysis for protein crystallization and structure determination. *Nature Methods, 4*, 1019-1021.

21. Derewenda, Z. S., & Vekilov, P. G. (2006). Entropy and surface engineering in protein crystallization. *Acta Crystallographica Section D, 62*, 116-124.
22. Kabsch, W. (2010). XDS. *Acta Crystallographica Section D: Biological Crystallography, 66*, 125-132.
23. Diederichs, K. (2017). Dissecting random and systematic differences between noisy composite data sets. *Acta Crystallographica Section D, Structural Biology, 73*, 286-293.
24. Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S., & Tucker, P. A. (2005). Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. *Acta Crystallographica Section D, 61*, 449-457.
25. CCP4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallographica Section D, 50*, 760-763.
26. Schneider, T. R., & Sheldrick, G. M. (2002). Substructure solution with SHELXD. *Acta Crystallographica Section D, 58*, 1772-1779.
27. Hao, Q. (2004). ABS: a program to determine absolute configuration and evaluate anomalous scatterer substructure. *Journal of Applied Crystallography, 37*, 498-499.
28. Terwilliger, T. C. (2000). Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr, 56*, 965-972.
29. Perrakis, A., Morris, R., & Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. *Nature Structural Biology, 6*, 458-463.
30. Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S., & Tucker, P. A. (2009). On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. *Acta Crystallographica Section D, 65*, 1089-1097.
31. Emsley, P., & Cowtan, K. (2004). Coot: model building tools for molecular graphics. *Acta Crystallographica Section D, Biological Crystallography, 60*, 2126-2132.
32. Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., & Adams, P. D. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D, Biological Crystallography, 68*, 352-367.
33. Chen, V. B., Arendall, W. B., III, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D, 66*, 12-21.
34. Krissinel, E., & Henrick, K. (2005). Detection of protein assemblies in crystals. In M. R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, & I. Fischer (Eds.), *Computational Life Sciences, Proceedings* (Vol. 3695, pp. 163-174).
35. Zhang, K. (2016). Gctf: Real-time CTF determination and correction. *Journal of Structural Biology, 193*, 1-12.
36. Punjani, A., Rubinstein, J. L., Fleet, D. J., & Brubaker, M. A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods, 14*, 290-296.
37. van Dijk, M., & Bonvin, A. M. J. J. (2009). 3D-DART: a DNA structure modelling server. *Nucleic Acids Research, 37*, W235-W239.
38. Kim, D. N., Moriarty, N. W., Kirmizialtin, S., Afonine, P. V., Poon, B., Sobolev, O. V., Adams, P. D., & Sanbonmatsu, K. (2019). Cryo_fit: Democratization of flexible fitting for cryo-EM. *Journal of Structural Biology, 208*, 1-6.
39. Brookes, E., Demeler, B., Rosano, C., & Rocco, M. (2010). The implementation of SOMO (SOlution MOdeller) in the UltraScan analytical ultracentrifugation data analysis suite: enhanced capabilities allow the reliable hydrodynamic modeling of virtually any kind of biomacromolecule. *European Biophysics Journal with Biophysics Letters, 39*, 423-435.
40. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soeding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology, 7*.
41. Robert, X., & Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research, 42*, W320-W324.
42. Laskowski, R. A., & Swindells, M. B. (2011). LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modelling, 51*, 2778-2786.
43. Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., Krasny, R., Wei, G. W., Holst, M. J., McCammon, J. A., & Baker, N. A. (2018). Improvements to the APBS biomolecular solvation software suite. *Protein Science, 27*, 112-128.

5.4 Chapter Summary

This chapter presents a thorough structural and functional investigation of the GntR-type NanR from *E. coli.* Solution studies in the presence of Neu5Ac were inconsistent with the literature as no Neu5Ac-induced dissociation was observed. The binding kinetics reported a nanomolar affinity for DNA, which was consistent with the literature and revealed novel cooperative behaviour. The stoichiometry of the NanR-DNA interaction was defined using multi-wavelength sedimentation velocity experiments, demonstrating that NanR binds as a dimer to a total of three direct GGTATA repeats that constitute the DNA recognition site, forming a hexameric NanR-DNA hetero-complex. A sequence analysis revealed a unique 32 residue N-terminal extension. Through mutagenesis this extension was hypothesised to facilitate the positive cooperativity, following observation of a perturbed ability to form high order hetero-complexes. Excitingly, the crystal structure of *E. coli* NanR was solved in complex with sialic acid and a zinc ion. This is the first known structure of a GntR-type sialoregulator to be reported, providing insight into the allosteric mechanism of regulation and novel evidence of a metal ion interacting with an effector molecule. Single particle cryo-EM experiments reported a structure of the NanR-DNA hetero-complex and provided structural evidence to support the multimeric hetero-complex assembly process. This is the first known structure of a NanR gene regulator in complex with DNA. Together, this characterisation of the GntR-type NanR from *E. coli* enhances our understanding of bacterial sialoregulation at a molecular level.

5.5 Chapter References

44. Reizer, A., Deutscher, J., Saier, M. H., Jr., & Reizer, J. (1991). Analysis of the gluconate (*gnt*) operon of *Bacillus subtilis*. *Molecular Microbiology, 5*, 1081-1089.
45. Rigali, S., Derouaux, A., Giannotta, F., & Dusart, J. (2002). Subdivision of the Helix-Turn-Helix GntR Family of Bacterial Regulators in the FadR, HutC, MocR, and YtrA Subfamilies. *Journal of Biological Chemistry, 277*, 12507-12515.
46. Jain, D. (2015). Allosteric control of transcription in GntR family of transcription regulators: A structural overview. *IUBMB Life, 67*, 556-563.
47. Suvorova, I. A., Korostelev, Y. D., & Gelfand, M. S. (2015). GntR Family of Bacterial Transcription Factors and Their DNA Binding Motifs: Structure, Positioning and Co-Evolution. *Public Library of Science One, 10*, e0132618-e0132618.
48. Brennan, R. G. (1993). The winged-helix DNA-binding motif - another helix-turn-helix takeoff. *Cell, 74*, 773-776.
49. Hoskisson, P. A., Rigali, S., Fowler, K., Findlay, K. C., & Buttner, M. J. (2006). DevA, a GntR-Like Transcriptional Regulator Required for Development in *Streptomyces coelicolor*. *Journal of Bacteriology, 188*, 5014.
50. Franco, I. S., Mota, L. J., Soares, C. M., & de Sa-Nogueira, I. (2007). Probing key DNA contacts in AraR-mediated transcriptional repression of the *Bacillus subtilis* arabinose regulon. *Nucleic Acids Research, 35*, 4755-4766.
51. Zheng, M., Cooper, D. R., Grossoehme, N. E., Yu, M., Hung, L. W., Cieslik, M., Derewenda, U., Lesley, S. A., Wilson, I. A., Giedroc, D. P., & Derewenda, Z. S. (2009). Structure of *Thermotoga maritima* TM0439: implications for the mechanism of bacterial GntR transcription regulators with Zn2+-binding FCD domains. *Acta Crystallographica, Section D, Biological Crystallography, 65*, 356-365.

52. Lee, M. H., Scherer, M., Rigali, S., & Golden, J. W. (2003). PlmA, a new member of the GntR family, has plasmid maintenance functions in *Anabaena* sp. strain PCC 7120. *Journal of Bacteriology, 185*, 4315-4325.
53. Bramucci, E., Milano, T., & Pascarella, S. (2011). Genomic distribution and heterogeneity of MocR-like transcriptional factors containing a domain belonging to the superfamily of the pyridoxal-5'-phosphate dependent enzymes of fold type I. *Biochemical and Biophysical Research Communications, 415*, 88-93.
54. DiRusso, C. C., Heimert, T. L., & Metzger, A. K. (1992). Characterization of FadR, a global transcriptional regulator of fatty acid metabolism in *Escherichia coli.* Interaction with the *fadB* promoter is prevented by long chain fatty acyl coenzyme A. *Journal of Biological Chemistry, 267*, 8685-8691.
55. van Aalten, D. M., DiRusso, C. C., Knudsen, J., & Wierenga, R. K. (2000). Crystal structure of FadR, a fatty acid-responsive transcription factor with a novel acyl coenzyme A-binding fold. *EMBO Journal, 19*, 5167-5177.
56. van Aalten, D. M., DiRusso, C. C., & Knudsen, J. (2001). The structural basis of acyl coenzyme A-dependent regulation of the transcription factor FadR. *EMBO Journal, 20*, 2041-2050.
57. Fujihashi, M., Nakatani, T., Hirooka, K., Matsuoka, H., Fujita, Y., & Miki, K. (2014). Structural characterization of a ligand-bound form of *Bacillus subtilis* FadR involved in the regulation of fatty acid degradation. *Proteins-Structure Function and Bioinformatics, 82*, 1301-1310.
58. Shi, W., Kovacikova, G., Lin, W., Taylor, R. K., Skorupski, K., & Kull, F. J. (2015). The 40-residue insertion in *Vibrio cholerae* FadR facilitates binding of an additional fatty acyl-CoA ligand. *Nature Communications, 6*, 6032.
59. Gao, Y. G., Suzuki, H., Itou, H., Zhou, Y., Tanaka, Y., Wachi, M., Watanabe, N., Tanaka, I., & Yao, M. (2008). Structural and functional characterization of the LldR from *Corynebacterium glutamicum*: a transcriptional repressor involved in L-lactate and sugar utilization. *Nucleic Acids Research, 36*, 7110-7123.
60. Jain, D., & Nair, D. T. (2013). Spacing between core recognition motifs determines relative orientation of AraR monomers on bipartite operators. *Nucleic Acids Research, 41*, 639-647.
61. Fillenberg, S. B., Grau, F. C., Seidel, G., & Muller, Y. A. (2015). Structural insight into operator dre-sites recognition and effector binding in the GntR/HutC transcription regulator NagR. *Nucleic Acids Research, 43*, 1283-1296.
62. Raman, N., Black, P. N., & DiRusso, C. C. (1997). Characterization of the fatty acid-responsive transcription factor FadR. Biochemical and genetic analyses of the native conformation and functional domains. *Journal of Biological Chemistry, 272*, 30645-30650.
63. Weickert, M. J., & Adhya, S. (1992). A family of bacterial regulators homologous to Gal and Lac repressors. *Journal of Biological Chemistry, 267*, 15869-15874.
64. Franco, I. S., Mota, L. J., Soares, C. M., & de Sa-Nogueira, I. (2006). Functional domains of the *Bacillus subtilis* transcription factor AraR and identification of amino acids important for nucleoprotein complex assembly and effector binding. *Journal of Bacteriology, 188*, 3024-3036.
65. Resch, M., Schiltz, E., Titgemeyer, F., & Muller, Y. A. (2010). Insight into the induction mechanism of the GntR/HutC bacterial transcription regulator YvoA. *Nucleic Acids Research, 38*, 2485-2497.
66. Saenger, W., Orth, P., Kisker, C., Hillen, W., & Hinrichs, W. (2000). The Tetracycline Repressor-A Paradigm for a Biological Switch. *Angewandte Chemie, 39*, 2042-2052.
67. Goeke, D., Kaspar, D., Stoeckle, C., Grubmuller, S., Berens, C., Klotzsche, M., & Hillen, W. (2012). Short peptides act as inducers, anti-inducers and corepressors of Tet repressor. *Journal of Molecular Biology, 416*, 33-45.
68. Kanwar, M., Wright, R. C., Date, A., Tullman, J., & Ostermeier, M. (2013). Protein switch engineering by domain insertion. *Methods Enzymology, 523*, 369-388.
69. Hardy, J. A., & Wells, J. A. (2004). Searching for new allosteric sites in enzymes. *Current Opinion In Structural Biology, 14*, 706-715.
70. Plumbridge, J., & Vimr, E. (1999). Convergent pathways for utilisation of the amino sugars *N*-acetylglucosamine, *N*-acetylmannosamine, and *N*-acetylneuraminic acid by *Escherichia coli*. *Journal of Bacteriology, 181*, 47-54.
71. Kalivoda, K. A., Steenbergen, S. M., Vimr, E. R., & Plumbridge, J. (2003). Regulation of sialic acid catabolism by the DNA binding protein NanR in *Escherichia coli*. *Journal of Bacteriology, 185*, 4806-4815.
72. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
73. Vimr, E. R. (2013). Unified theory of bacterial sialometabolism: how and why bacteria metabolise host sialic acids. *International Scholarly Research Notices in Microbiology, 2013*, 816713.
74. Vimr, E. R., Kalivoda, K. A., Deszo, E. L., & Steenbergen, S. M. (2004). Diversity of microbial sialic acid metabolism. *Microbiology & Molecular Biology Reviews, 68*, 132-153.

75. Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res., 14*, 1188-1190.
76. Schuck, P. (2007). *Protein interactions: biophysical approaches for the study of complex reversible systems* (Vol. 5). New York, NY, Springer.
77. Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols, 2*, 1849-1861.
78. Westphal, L. L., Sauvey, P., Champion, M. M., Ehrenreich, I. M., & Finkel, S. E. (2016). Genomewide Dam Methylation in *Escherichia coli* during Long-Term Stationary Phase. *mSystems, 1*.
79. Chu, D., Roobol, J., & Blomfield, I. C. (2008). A theoretical interpretation of the transient sialic acid toxicity of a *nanR* mutant of *Escherichia coli*. *Journal of Molecular Biology, 375*, 875-889.

Chapter Six

Defining the DNA-binding mechanism of the GntR-type sialoregulator from *Escherichia coli* – Part Two

6.1 Chapter overview

This chapter provides additional and supporting information to Chapter Five. Combined, the overall aim of this body of work is to develop a mechanistic understanding of how the GntR-type NanR from *E. coli* regulates gene expression of sialic acid catabolism.

The thermal stability of *E. coli* NanR is investigated using nano-differential scanning fluorimetry experiments and confirms an increase in thermal stability in the presence of sialic acid and its DNA recognition sequence. In addition, electrophoretic mobility shift assays (EMSA) investigate the effect on DNA-binding activity when the spacing between the GGTATA binding motif is increased. A thorough investigation into the stoichiometry of the protein-DNA complex is presented using single-wavelength sedimentation velocity experiments. These experiments provide the foundation for the multi-wavelength sedimentation velocity analysis presented in Chapter Five. As this is an emerging strategy, the data analysis of multi-wavelength sedimentation velocity experiments is detailed in this chapter. The C-terminal domain sub-structure used to solve the X-ray structure of *E. coli* NanR using a combined molecular replacement and single wavelength anomalous dispersion strategy is presented. Although crystallisation trials conducted in the presence of DNA were not suitable for structure determination, the crystals obtained along with strategies undertaken during optimisation are described.

6.2 Results and Discussion

6.2.1 The thermal stability of *E. coli* NanR

To investigate the thermal stability of *E. coli* NanR, differential scanning fluorimetry (DSF) experiments were employed. Initially, these experiments were conducted using SYPRO Orange, a fluorescent dye that binds hydrophobic residues and reports during protein unfolding (1). However, a high background fluorescence was observed from the start of the melt curve, despite using a low final concentration of SYPRO Orange (as per the manufacturer's instructions). This indicated that the fluorescent dye was binding the protein before it has unfolded. To investigate this, circular dichroism (CD) spectroscopy was performed to check the protein was folded in solution (Figure 6.1A) as described in Chapter Eight, Section 8.6.3.2.

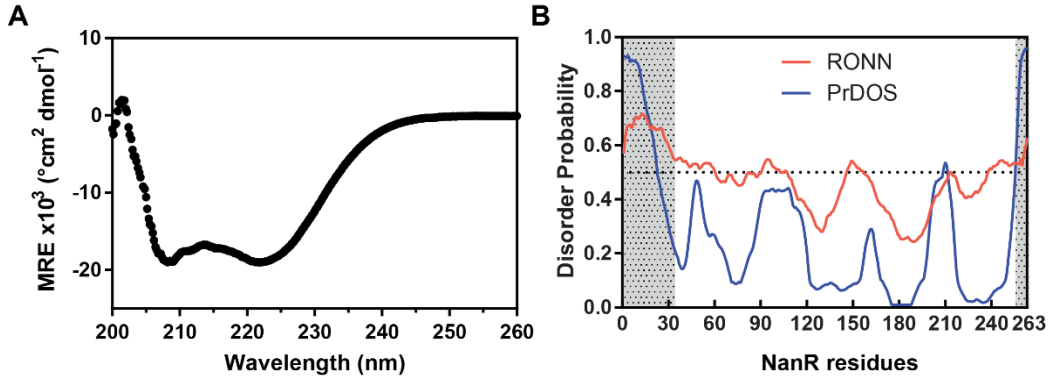


Figure 6.1| CD spectra and disorder plot of *E. coli* NanR. A) The negative peaks at 208 nm and 222 nm are characteristic of α-helices and suggests that NanR is folded in solution. Experiment was conducted using 0.05 mg mL^{-1} NanR in 20 mM Tris-HCl (pH 8.0) using a 1 cm pathlength quartz cuvette. The data was converted to mean residue ellipticity (MRE) and plotted as a function of wavelength (nm). **B)** The disorder probability estimated from RONN (https://www.strubi.ox.ac.uk/RONN) and PrDOS (http://prdos.hgc.jp/cgi-bin/top.cgi) online servers. The regions of disorder are highlighted (grey with dot).

The CD spectra for NanR displayed negative peaks at 208 nm and 222 nm, which is characteristic for a protein comprised of α-helices (2). This observation was consistent with the crystal structure presented in Chapter Five, where the C-terminal domain displayed an all α-helical architecture. Taken together, this supports that NanR is folded in solution. Consequently, the cause of the high background fluorescence observed when using SYPRO® Orange is likely attributed to the 32 residue N-terminal extension of the DNA binding domain, which was predicted to be disordered (Figure 6.1B) by the online servers RONN and PrDOS. In addition, its composition is primarily comprised of hydrophobic residues (11, 34%). Thus, the hydrophobic dye SYPRO® is not suitable to investigate the thermal stability of *E. coli* NanR.

As an alternative strategy, the Prometheus NT.48 instrument (NanoTemper Technologies) was employed to determine the thermal stability of NanR without the use of environmentally sensitive fluorescent dyes. This was achieved by monitoring the intrinsic fluorescence of tryptophan residues upon introduction of a thermal gradient. Here, as the tertiary structure of the protein unfolds, the intrinsic tryptophan fluorescence changes as the chemical environment of these residues is altered (3). These changes are monitored by dual-UV detection at emission wavelengths of 350/330 nm, where the ratio of these signals can be used to infer the unfolding temperature of the protein with high resolution (4). Furthermore, these experiments can be conducted across any buffer system and require a sample volume of only 10 µL. Thus, this technology is referred to as Nano-DSF.

Nano-DSF experiments were carried out using a NanR concentration of 1 mg mL^{-1} in 20 mM Tris-HCl (pH 8.0), 150 mM NaCl. Here, NanR displayed biphasic unfolding behaviour with the identification of two unfolding transitions in the first derivative plot (Figure 6.2, blue). The first transition (T_m^1) gave a melting temperature of 52.04 ± 0.01 °C, with an onset melting temperature (T_{onset}) of 48.8 ± 0.04 °C (Table 6.1). The second transition (T_m^2) although visible (~67 °C), was not distinguished as an inflection point in the PR.ThermControl software (NanoTemper Technologies) so no melting temperature was recorded. This is likely due to only a minor difference in the 350/330 nm emission wavelengths. When supplemented with Neu5Ac, a similar biphasic unfolding was observed (Figure 6.2, red), where the T_m^1 gave a melting temperature of 53.97 ± 0.08 °C, with a T_{onset} of 50.04 ± 0.82 °C, while T_m^2 gave a melting temperature of 68.1 ± 0.1 °C (Table 6.1). The height of these transition peaks suggested almost a 50:50 population of species unfolding at T_m^1 versus T_m^2. Conversely, when NanR was supplemented with its DNA recognition site, this biphasic unfolding was lost, as only one transition peak was observed. This transition gave a melting temperature of 69.9 ± 0.01 °C and a T_{onset} of 61.41 ± 0.07 °C (Table 6.1).

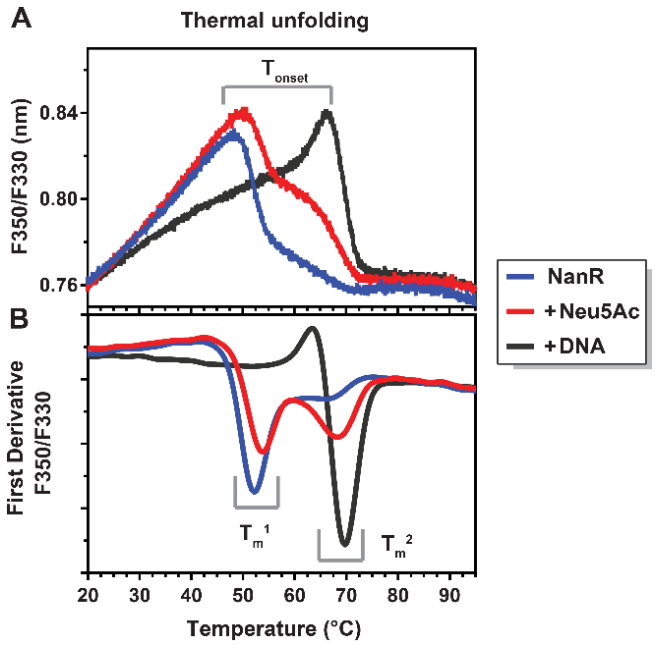


Figure 6.2| Thermal stability of *E. coli* NanR using nano-DSF. A) The change in the ratio of intrinsic tryptophan fluorescence at the emission wavelengths of 350 and 330 nm. The onset melting temperature (T_{onset}) is highlighted. **B)** The first derivative of the fluorescence ratio curve. The peak maximum corresponds to the melting temperature (T_m) of the protein/ transition, derived from the inflection point of the ratio curve above. Transitions T_m^1 and T_m2 are highlighted. Nano-DSF experiments were performed with 1 mg mL^{-1} (33 µM) protein in 20mM Tris-HCl (pH 8.0), 150 mM NaCl and additionally supplemented with either 10 mM Neu5Ac or 300 µM DNA. Thermal unfolding was monitored from 20 to 95 °C at a at a ramp rate of 1 °C per min for NanR (blue), NanR + Neu5Ac (red) and NanR + DNA (black).

Interestingly, all samples displayed a decrease in the fluorescent ratio (350/330 nm), which is consistent with a blue shift (Figure 6.2A). This feature indicates that the tryptophan residues are solvent exposed and begin to accommodate a more hydrophobic environment upon heating. Conversely, an increase in the fluorescent ratio would be consistent with a red shift, where a buried tryptophan residue slowly becomes solvent exposed upon heating (5). In the structure of *E. coli* NanR, the two tryptophan residues (residues 198 and 252) are located near the dimer interface and towards the C-terminus, respectively. Both these residues are solvent exposed, supporting the observed blue shift.

Table 6.1 | Summary of data analysis from nano-DSF experiments. The T_{onset} and T_m values represent averages along with the standard deviation from duplicates.

Sample	T_{onset} (°C)	T_m^1 (°C)	T_m^2 (°C)
NanR	48.8 ± 0.04	52.04 ± 0.01	-
NanR + Neu5Ac	50.04 ± 0.82	53.97 ± 0.08	68.1 ± 0.1
NanR + DNA	61.41 ± 0.07	-	69.9 ± 0.01

Combined, these data demonstrated that NanR displays biphasic unfolding in the presence and absence of Neu5Ac, although this behaviour was altered in the presence of DNA. However, all data exhibited a blue shift with a decrease in the fluorescent ratio (350/330 nm). This feature suggests that the first transition (T_m^1) observed in the presence and absence of Neu5Ac corresponds to a partial denaturation of the protein, while the second transition (T_m^2) observed in all samples corresponds to complete denaturation of the protein. Although, a second interpretation would be that each transition reflects the two domains of NanR unfolding at different temperatures.

To further understand this thermal denaturation process, it is important to consider NanR is comprised of two domains; an N-terminal DNA-binding domain and a C-terminal effector-binding domain. Interestingly, when in the presence of Neu5Ac, both melting temperatures (T_m^1 and T_m^2), as well as the onset melting temperature had increased by approximately 2 °C relative to NanR alone—this is consistent with an increase in thermal stability. Furthermore, the second transition peak was more pronounced in the presence of Neu5Ac compared to NanR alone (Figure 6.2A and B, respectively in red). This suggests that the second transition represents denaturation of the C-terminal domain, where the increase in signal with Neu5Ac is consistent with effector binding, while the first transition represents denaturation of the N-terminal domain. Conversely, in the presence of DNA, only a single melting temperature was measured, with a significant increase in the onset melting temperature by 10-12 °C. This demonstrates that when bound to DNA, NanR is sustainably more stable in solution.

6.2.2 Analysing the distance between GGTATA binding repeats using electrophoretic mobility shift assays

To achieve regulation, *E. coli* NanR binds a conserved operator site that contains three direct repeats of the DNA sequence GGTATA. This results in the formation of three concentration-dependent complexes or discrete bands by EMSA (Figure 6.2A). Separating these direct binding repeats are the base pairs ACA between direct repeats one and two (spacer region 1), and base pairs AA between direct repeats two and three (spacer region 2). Collectively, these base pairs are conserved in the promoter region of the three operons NanR regulates as part of the sialoregulon (6). However, the biological relevance of maintaining a third GGTATA repeat is currently unresolved and thus an open question. To address this and investigate whether spacing of the repeats is crucial for DNA binding, the number of base pairs was progressively increased by three and subsequently assessed using EMSA. A total of six mutant DNA oligonucleotides were designed and tested below, each with an increasingly larger distance between the GGTATA direct repeat (Figure 5.2B-G). An identical titration range for NanR and a constant concentration of oligonucleotide is used is each gel shift assay.

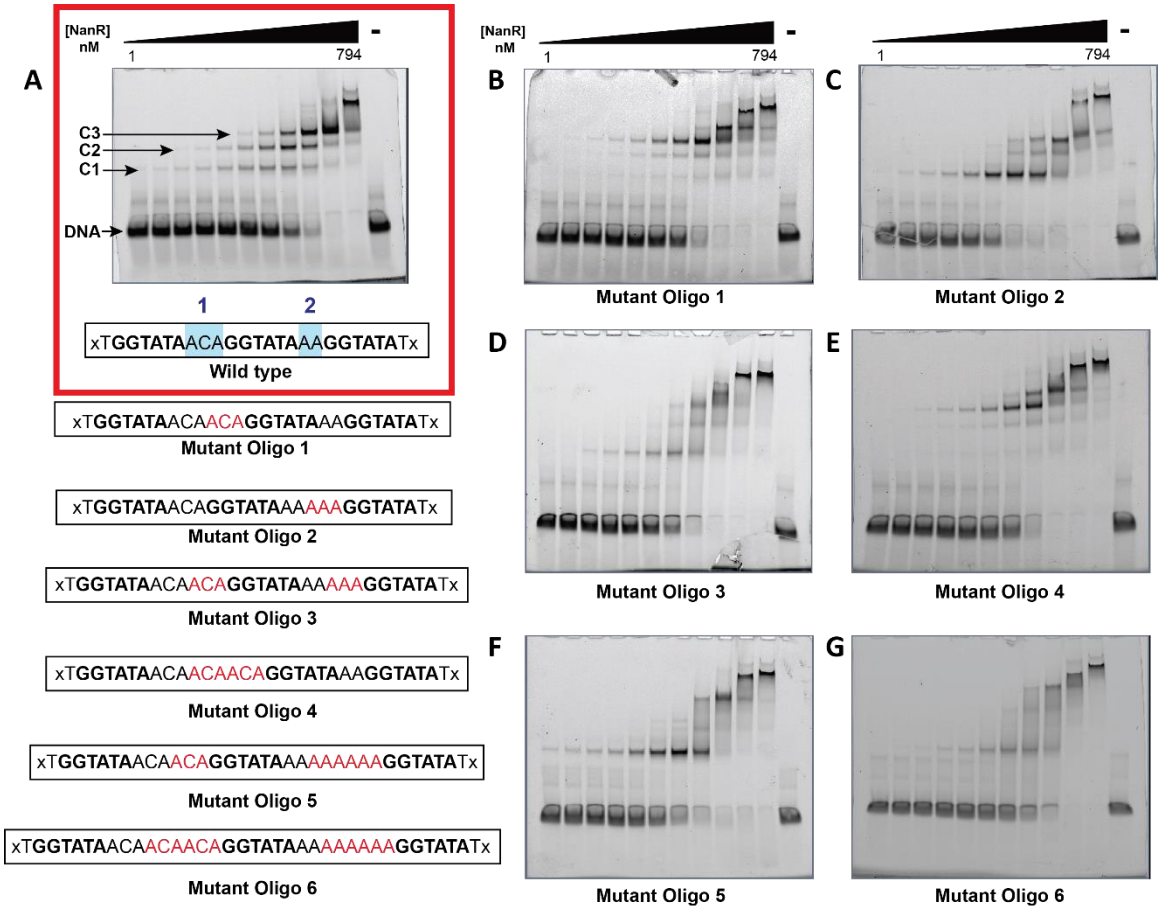


Figure 6.2 | Analysing the distance between binding repeats using EMSA. A) Wild-type sequence. Three concentration dependent complexes are visualised, labelled C1, C2 and C3. The DNA probe is labelled DNA. The spacer region between GGTATA repeats one and two is labelled 1, while the spacer region between GGTATA repeats two and three is labelled 2. Both are highlighted blue. **B to G)** EMSA gel for mutant oligonucleotides 1-6. The left side of the figure presents the DNA sequence for each mutant. Highlighted in red is the addition base pairs within the spacer.

It is evident that the distance between spacer region 2 is significantly more important for the assembly of higher order complexes, than the distance between spacer region 1. In mutant oligonucleotide one, where the distance had increased by three base pairs, the appearance of a hexameric NanR-DNA hetero-complex species was still observed, with a retarded band at the same height as C3 in the wild type gel (Figure 6.2B). Conversely, the data for mutant oligonucleotide two primarily demonstrated hetero-complex formation that was consistent with C1, or a dimeric NanR-DNA hetero-complex species. There was little appearance of bands that were consistent with a higher-order hetero-complex, suggesting that the DNA-binding affinity was reduced with this nucleotide sequence (Figure 6.2C). This supported the notion that the third GGTATA repeat is important to promote the cooperative assembly of the hexameric NanR-DNA hetero-complex. In mutant oligonucleotide three, both spacer regions have been increased by three base pairs, where a significant loss in binding affinity is observed, and a further

decrease in signal for the retarded bands that correspond to complex formation (Figure 6.2C). This is consistent with a combined effect of what is observed in mutant oligonucleotide one and two. Mutant oligonucleotide four and five have been increased by a further three base pairs at spacer region 1 and 2, respectively. In these gel shift assays, a similar assessment can be made as to what was observed in Figure 6.2A and 6.2B. This observation is similar in mutant oligonucleotide five, which shows the DNA probe being retarded just as quickly (Figure 6.2F). Lastly, when the distance between the GGTATA repeats was increased to a total of six base pairs, very little complex formation was visualised (Figure 6.2G). While the last or most right-hand lane in each gel was a DNA only control, the inside two to three lanes showed a significant shift with respect to the DNA probe. Moreover, these lanes appeared more smeared than others. As this was also a feature in the wild type, these bands are likely the result of non-specific binding leading to a heterogeneous population of protein-DNA intermediates that would fail to produce a discrete shift.

Combined, these mutated oligonucleotides explored what effect distance has on DNA-binding affinity for NanR. Here, it was discovered that when this distance was increased between spacer region 2, as opposed to the first spacer region, a significant loss in DNA binding affinity can be observed. This demonstrates that the third GGTATA repeat is crucial for the assembling the higher order hetero-complex. Interestingly, this data also shows that the pairs of direct GGTATA repeats are not equivalent, suggesting that cooperativity of the system is mediated by only one of these pairs. As higher order hetero-complex assembly was perturbed when spacer region 2 was mutated, perhaps GGTATA repeats 2 and 3 mediate the cooperative behaviour of the system. However, it is unclear if the binding of NanR occurs sequentially to these repeats or through a random ordered process.

6.2.3 Sedimentation velocity analysis on the *E. coli* NanR-DNA interaction

Prior to this study, the stoichiometry of the *E. coli* NanR-DNA interaction was poorly understood. This is because the gel-based technique used to estimate the size of macromolecules gave largely inconsistent results. However, in Chapter Five, analytical ultracentrifugation was employed to quantitatively demonstrate that *E. coli* NanR coordinates its DNA recognition as dimers, forming a hexameric NanR-DNA hetero-complex. This was achieved using novel multi-wavelength sedimentation velocity experiments, conducted at the University of Lethbridge, Canada as this advanced analytical ultracentrifugation instrumentation was not available in New Zealand.

However, prior to these experiments, single-wavelength sedimentation velocity experiments were initially performed using a titration range of *E. coli* NanR against a fixed concentration of fluorescently end-labelled DNA, as described in Chapter Eight (*Section 8.6.3.4*). Here, by measuring the emission

wavelength of the fluorescein tag at 495 nm, I could detect free DNA, and evaluate protein-DNA complex formation, indicated by a positive shift in the sedimentation coefficient. Furthermore, in order to accurately determine the molecular weight of the species in solution, the experiment was repeated at two velocities as this enables the hydrodynamic information to be extracted with increased precision.

6.2.3.1 Single-wavelength sedimentation velocity analysis using two GGTATA binding sites

Initially, experiments were conducted with an oligonucleotide containing only two GGTATA binding sites in order to simplify the system, as by EMSA only two concentration-dependent hetero-complexes are visualised (Figure 6.3A).

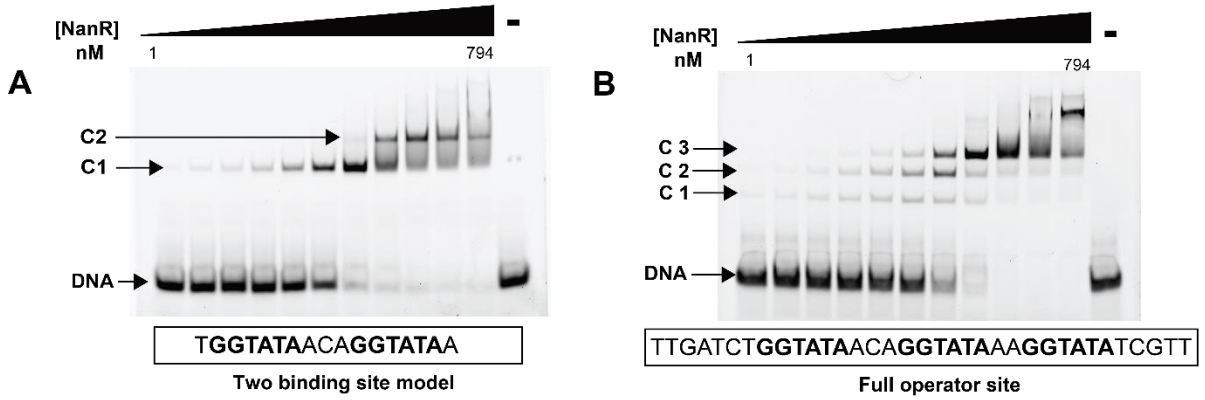


Figure 6.3 | EMSA gel of the two binding site model for *E. coli* NanR. A) When only two GGTATA repeats are present, NanR forms two concentration-dependent hetero-complexes, labelled C1 and C2. **B)** When three GGTATA repeats are present, NanR forms three concentration-dependent hetero-complexes, labelled C1-3. The free DNA probe for each is labelled and the DNA sequence is shown below each respective EMSA gel.

In this experiment, a total of seven individual samples were collected at 60 000 rpm and 28 000 rpm, where one was a DNA only control, while the remaining six samples included a micromolar titration range of *E. coli* NanR. Following data collection at each speed, the sedimentation data was evaluated using a two-dimensional spectrum analysis (2DSA) (7) and further refined by 50 Monte Carlo iterations using the state-of-the-art software, Ultrascan (8). Here, upon increasing concentration two new peaks could be observed, both displaying a larger sedimentation coefficient relative to the DNA only control of 2.19 S (Figure 6.4A). In addition, the free DNA signal slowly decreased upon increasing concentration of NanR. Combined, these observations were consistent with the formation of a protein-DNA hetero-complex of two different sizes. Interestingly, the small peak at a sedimentation coefficient of approximately 1 S, which was consistent with a low concentration of single-stranded DNA, did not change during the NanR titration. This showed that only the double-stranded DNA interacted with and bound NanR.

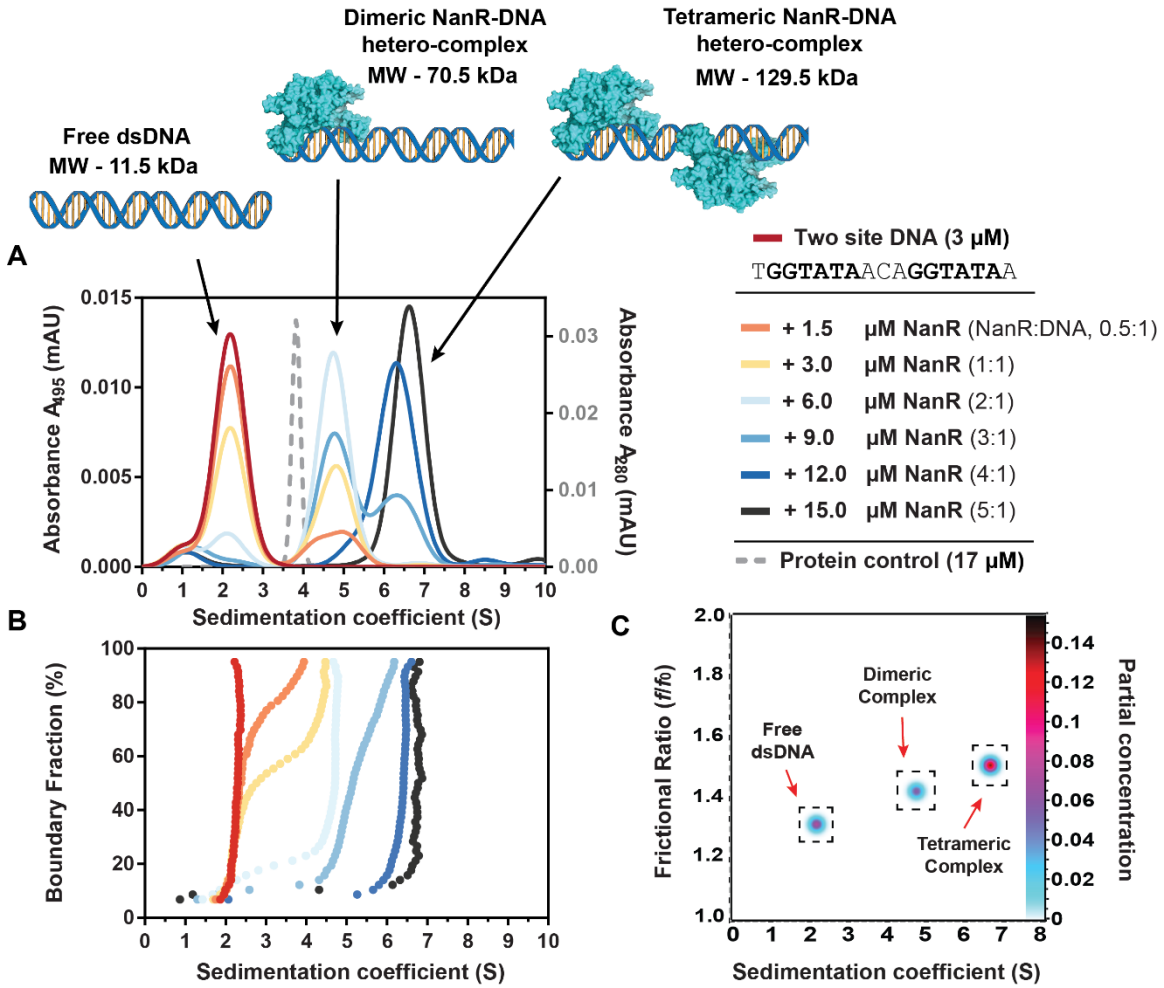


Figure 6.4 | Single-wavelength sedimentation velocity analysis of *E. coli* NanR and two GGTATA binding sites. A) 2DSA-Monte Carlo model displaying the sedimentation coefficient (S) distribution from the 60 000 rpm experiment. Each colour represents a different titration sample, collected independently (see legend). To illustrate the identity of the three main species in solution, a cartoon depiction of DNA and each NanR-DNA hetero-complex is presented. For clarity the sedimentation coefficient distribution for *E. coli* NanR collected independently is shown. B) van Holde-Weischet plot provides a graphical transformation of the sedimentation data allowing the proportional contributions of each species to be interpreted. C) Pseudo-3D plot presenting the global solute distribution, across all sedimentation data shown in (A) as a function of frictional ratio and sedimentation coefficient. Highlighted are the three predominant species in solution.

Table 6.2 | Summary of single-wavelength sedimentation velocity data collection and analysis from *E. coli* NanR

Data collection parameters						
Buffer density (g cm^3)					1.0059	
Buffer viscosity (cp)					1.0211	
Partial specific volume (mL g^{-1})						
E. coli NanR					0.7327	
DNA oligonucleotide (Two binding site model)					0.5550	
Dimeric NanR-DNA hetero-complex					0.7030	
Tetrameric NanR-DNA hetero-complex					0.7105	
Sample	**Sedimentation Coefficient (x10^{-13} S)**	**Diffusion Coefficient (x10^{-07} D)**	**Measured Molar Mass (kDa)**	**Oligomeric State of Complex**	**Expected Molar mass (kDa)**	**% Error (Molar mass)**
DNA only	2.19	10.34	11	-	11.5	1
0.5:1*	4.75	6.06	64	Dimeric	70.5	10
1:1	4.73	5.15	75	Dimeric	70.5	6
2:1	4.73	5.26	73	Dimeric	70.5	4
3:1	4.77	5.54	71	Dimeric	70.5	1
4:1	6.23	4.23	127	Tetrameric	129.5	2
5:1	6.61	4.33	131	Tetrameric	129.5	1

*Protein:DNA ratio

Following peak integration of these species, the measured molar mass of the complex at approximately 4.7 S (reported in Table 6.2) was consistent with the theoretical molar mass of a single NanR dimer and its DNA recognition sequence forming a dimeric NanR-DNA hetero-complex (70.5 kDa). Conversely, the measured molar mass of the hetero-complex at approximately 6.4 S (reported in Table 6.2) was consistent with the theoretical molar mass of two NanR dimers and its DNA recognition sequence forming a tetrameric NanR-DNA hetero-complex (129.5 kDa). Validation of this stoichiometry was supported by the small difference in molar mass between these values with an error less than 10% (Table 6.2). The accuracy of these measured molar masses is increased when a stable species is observed, as opposed to an equilibrium between two species that lie within a reaction boundary. This can be observed in the van Holde-Weischet plot (Figure 6.4B), which provides a graphical transformation of the sedimentation data allowing the proportional contributions of each species to be interpreted. A stable species can be visualised by a predominately straight line, which indicates the sample is primarily homogenous. For example the 2:1 (protein:DNA) sample is primarily a single species (hetero-complex), with a small amount of DNA. This distribution was also observed in the multi-wavelength data in Chapter Five and further reinforces why this protein-DNA ratio was selected for single-particle cryo-EM experiments in order to investigate the DNA bound conformation of NanR. In addition, the van Holde-Weischet plot provides insight into the kinetics of the NanR-DNA hetero-complex assembly process. In this study, the formation of the dimeric NanR-DNA hetero-complex appeared much slower, while the transition to the higher order hetero-complex or tetrameric NanR-DNA hetero-complex was much faster (Figure 6.4B). This difference in DNA binding kinetics was also

observed by EMSA (Figure 6.3A) and can be supported by the cooperative behaviour of the system, which was demonstrated in Chapter Five.

Furthermore, these species were combined to create a global model and then further refined by utilising genetic algorithms as part of the UltraScan software. This strategy removed regions with very low signal (false positives) from the model, which were generated by stochastic noise in the data and in turn emphasised the signal of the primary solutes or components that were present in solution (9). The result of this global refinement is presented as a Pseudo-3D plot, which allows hydrodynamic parameters such as the sedimentation coefficient, diffusion coefficient or frictional ratio to be compared. Figure 6.4C presents this plot as a function of frictional ratio and sedimentation coefficient, highlighting the free DNA, dimeric and tetrameric NanR-DNA hetero-complexes. Here, an increasing trend in the frictional ratio was observed as the sedimentation coefficient become larger. Since the frictional ratio is a measure of shape asymmetry, this feature was expected, since during complex assembly, the addition of successive NanR dimers would create an increasingly asymmetric molecule.

Taken together, these sedimentation velocity experiments demonstrated that when two GGTATA binding sites were present, two protein-DNA hetero-complexes of different sizes were formed. Further analysis defined the stoichiometry of these protein-DNA complexes, as a dimeric and tetrameric NanR-DNA hetero-complex, respectively. This was supported by the small difference between measured and theoretical molar mass values. In addition, van Holde-Weischet analysis illustrated the difference in DNA binding kinetics during this hetero-complex assembly process.

6.2.3.2 Single-wavelength sedimentation velocity analysis using three GGTATA binding sites

Next, to investigate the assembly process with the full operator site, sedimentation velocity experiments were repeated using three GGTATA binding sites, which produces three concentration-dependent complexes in EMSA (Figure 6.3B). In this experiment the initial velocities were adjusted to 50 000 rpm and 32 000 rpm since larger species were expected than previously, owing to the extra GGTATA repeat. However, the data was analysed and refined using UltraScan (8) in an analogous process. The 2DSA-Monte Carlo model displaying the sedimentation coefficient distribution from the 50 000 rpm experiment is presented below in Figure 6.5, while all data analysis is summarised in Table 6.3.

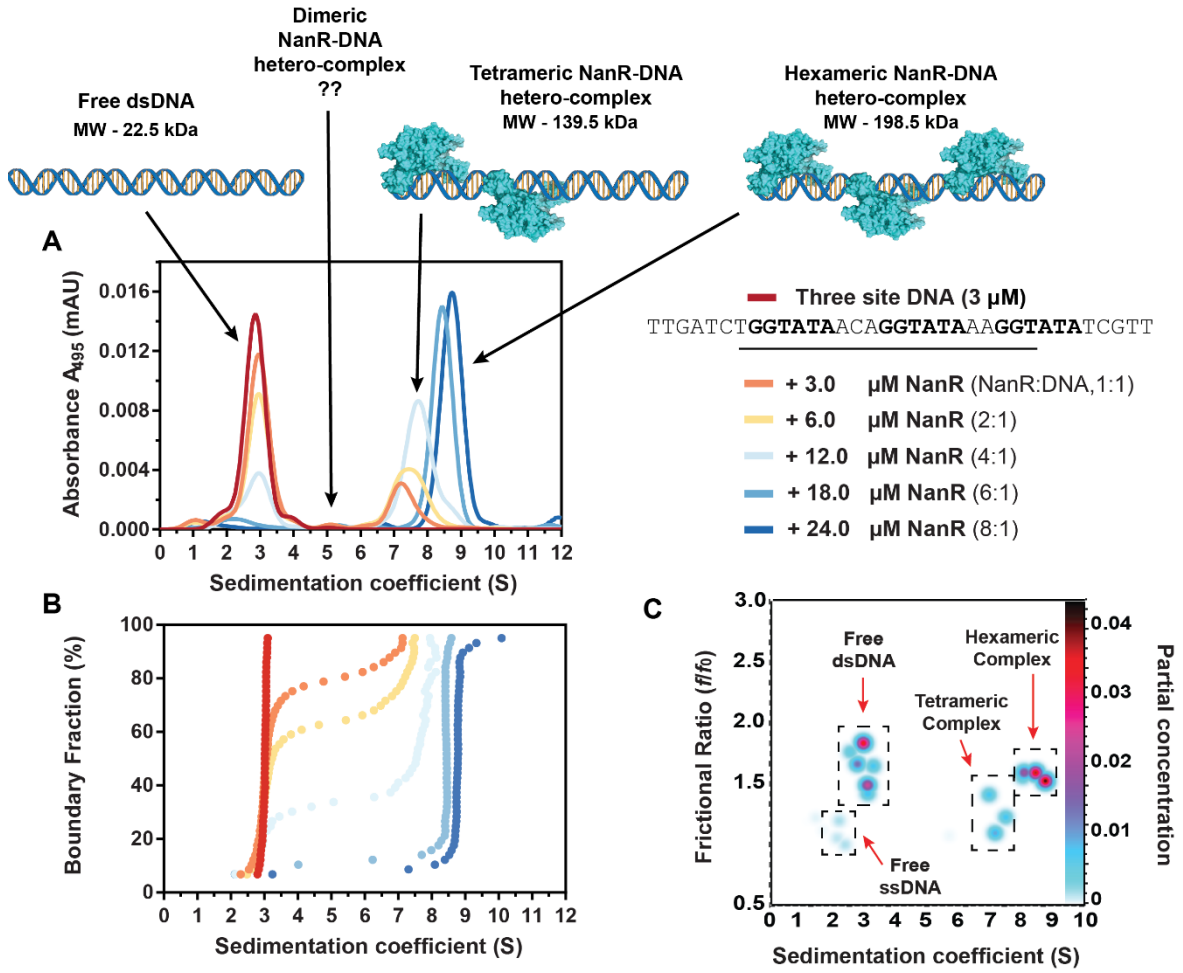


Figure 6.5 | Single-wavelength sedimentation velocity analysis of *E. coli* NanR and three GGTATA binding sites. A) 2DSA-Monte Carlo model displaying the sedimentation coefficient (S) distribution from the 50 000 rpm experiment. Each colour represents a different titration sample, collected independently (see legend). To illustrate the identity of the three main species in solution, a cartoon depiction of DNA and each NanR-DNA hetero-complex is presented. **B)** van Holde-Weischet plot provides a graphical transformation of the sedimentation data allowing the proportional contributions of each species to be interpreted. **C)** Pseudo-3D plot presenting the global solute distribution, across all sedimentation data shown in (A) as a function of frictional ratio and sedimentation coefficient. Highlighted are the three predominant species in solution.

In this experiment, upon increasing concentration of NanR, two prominent shifts in sedimentation coefficient were observed (Figure 6.5A). This was consistent with protein-DNA hetero-complex formation *via* a multistep assembly process. Following peak integration of these species, the measured molar mass of the peak centred at approximately 3 S (reported in Table 6.3) was consistent with the theoretical molar mass of the full operator site with three GGTATA binding sites (22.5 kDa). This small, yet positive shift in sedimentation coefficient relative to the two binding site model, is a result of the additional 18 base pairs, which includes the third GGTATA repeat (Figure 6.3). Upon titration of NanR, this peak slowly decreased and initially shifted to a peak centred at a sedimentation coefficient of

approximately 7 S. The measured molar mass of this species was consistent with the theoretical molar mass of two NanR dimers and its DNA recognition sequence forming a tetrameric NanR-DNA hetero-complex (139.5 kDa). Upon increasing the concentration further, this peak shifted to a peak centred at a sedimentation coefficient of approximately 8.6 S. Following peak integration, the measured molar mass of this species was consistent with the theoretical molar mass of three NanR dimers and its DNA recognition sequence forming a hexameric NanR-DNA hetero-complex (198.5 kDa). Having formed this high-order assembly, the system appears to saturate as no free DNA remains to be complexed. The proportional contributions of each species during this multistep assembly process can be visualised in the van Holde-Weischet plot (Figure 6.5B). In addition, the frictional ratio of these species can be visualised in the Pseudo-3D plot following further refinement with a genetic algorithm in UltraScan (Figure 6.5C).

Table 6.3 | Summary of single-wavelength sedimentation velocity data collection and analysis from *E. coli* NanR

Data collection parameters						
Buffer density (g cm^3)					1.0059	
Buffer viscosity (cp)					1.0211	
Partial specific volume (mL g^{-1})						
E. coli NanR					0.7327	
DNA oligonucleotide (Full operator site)					0.5550	
Tetrameric NanR-DNA hetero-complex					0.7034	
Hexameric NanR-DNA hetero-complex					0.7120	
Sample	**Sedimentation Coefficient ($x10^{-13}$ S)**	**Diffusion Coefficient ($x10^{-07}$ D)**	**Measured Molar Mass (kDa)**	**Oligomeric State of Complex**	**Expected Molar mass (kDa)**	**% Error (Molar mass)**
DNA only	2.97	6.93	23	-	22.5	1
1:1*	7.46	4.33	142	Tetrameric	139.5	2
2:1	7.51	4.21	146	Tetrameric	139.5	4
4:1	7.70	4.02	156	Tetrameric	139.5	11
6:1	8.49	3.39	211	Hexameric	198.5	6
8:1	8.77	3.36	219	Hexameric	198.5	10

*Protein:DNA ratio

Interestingly, in comparison to the two binding site model, no prominent peak was visible at a sedimentation coefficient of 5 S, which would be consistent with the formation of a dimeric NanR-DNA hetero-complex. However, as a tetrameric and hexameric NanR-DNA hetero-complex was observed, this highlighted that when the full promoter site is present, NanR can rapidly associate to form the higher order hetero-complexes. Thus, despite their presence in the multistep assembly process the signal concentration for the dimeric NanR-DNA hetero-complex was so low that it fell outside the detectable range for this experiment. This feature was also visualised by EMSA (Figure 6.3B). Furthermore, this assembly process was consistent with a cooperative system, as demonstrated in Chapter Five.

6.2.3.3 An introduction to multi-wavelength sedimentation velocity analysis

The data from the single-wavelength sedimentation velocity experiments presented above demonstrate the stoichiometry of the NanR-DNA hetero-complex. However, to provide a more definite answer to this complex system I employed a new analytical ultracentrifugation strategy, which is denoted multi-wavelength sedimentation velocity analysis. This emerging strategy offers a significant improvement in resolution by coupling a novel multi-wavelength detector (10) with high-performance computing. Furthermore, data collection involves an additional spectral dimension, providing the ability to distinguish between dissimilar macromolecules (such as protein and DNA) by exploiting the differences in their respective absorbance profiles (10; 11). Overall, this feature provides the capacity to characterise the assembly process of interacting components within a complex mixture, without any detection labels. The steps necessary to facilitate data analysis of multi-wavelength experiments are summarised below.

Initially, an absorbance profile for each interacting component must be collected using a dilution series over a range of wavelengths. This range must be tailored for each experiment based upon the spectral properties of the interacting components. In this study, a wavelength range of 220-320 nm was chosen as this encompasses the spectral profiles of both protein and DNA. Next, using an extinction coefficient of each interacting component, the dilution series is globally fitted in the software UltraScan to obtain an intrinsic extinction profile for the entire wavelength range measured (12). This profile is later used to scale the sedimentation data to molar concentration and subsequently decompose the absorbance signal of the interacting components into separate profiles. However, in order to accurately achieve spectral decomposition, two key factors must be assured, including: 1) an accurate extinction coefficient when fitting the intrinsic extinction profile; and 2) the intrinsic extinction profile of the individual components must be sufficiently different or 'orthogonal'. To determine orthogonality, the vector angle between the two spectral profiles can be measured using UltraScan, where an angle of 90 ° reflects no spectral overlap and a vector angle of 0 ° indicates perfect overlap. In this study, the spectral profiles for *E. coli* NanR and DNA, presented in Figure 6.6, gave a vector angle of 60 °. This indicated that they were sufficiently different and could be considered 'orthogonal'.

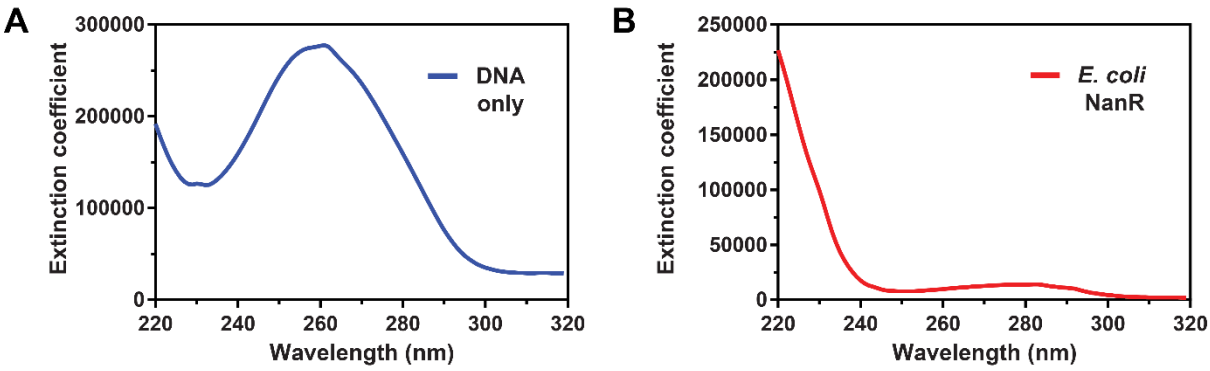


Figure 6.6 | Intrinsic extinction profiles for the interacting partners. A) Intrinsic extinction profiles for *E. coli* NanR **B)** Intrinsic extinction profiles for DNA (Two binding site model). Together, these profiles provide an extinction coefficient for the entire wavelength range measured. A vector angle of 60 ° between these profiles, supports that they are 'orthogonal'—this feature is essential to achieve spectral decomposition (12).

In the next step, sedimentation velocity data is collected across the wavelength range of interest within a two-sector cell. In Chapter Five, protein-DNA titrations of 1:1, 2:1, 4:1, 5:1, 6:1 and 10:1 (μM ratios) were employed to simulate and further expand the single-wavelength experiments above. Moreover, sedimentation velocity data were collected individually on purified NanR and pure DNA using a wavelength of 280 nm and 260 nm, respectively. Together, these serve as the controls for the experiment, thus any observation of a positive shift in sedimentation coefficient supports an interaction between NanR and DNA.

By introducing the spectral dimension (or alternatively, a fourth dimension) to this new approach, the volume of data over traditional methods is significantly increased when combined with data collection over a range of wavelengths (10). As a result, multi-wavelength sedimentation velocity experiments offer a substantial improvement in spectral resolution over single-wavelength approaches (12), which comprise a three-dimensional (3D) dataset that includes time, radius and absorbance (13). An example of this new multi-wavelength sedimentation velocity data is presented in Figure 6.7. Following data collection, each wavelength can be analysed independently. While this process is computationally demanding, the analysis is manageable *via* access to high-performance computing and use of the software UltraScan, which evaluates the sedimentation data using an iterative 2DSA method (7).

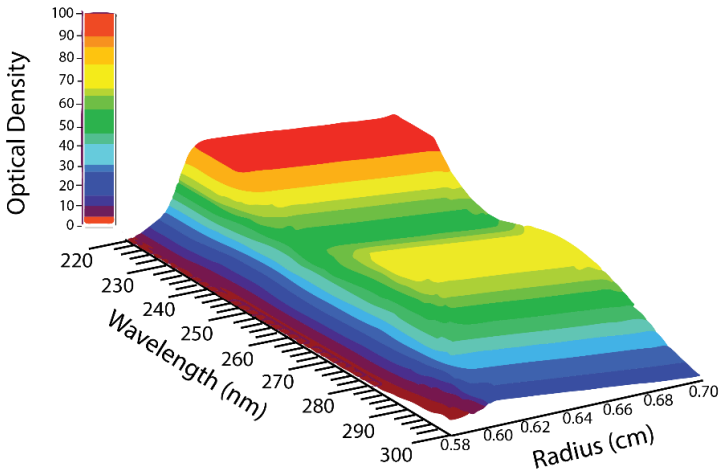


Figure 6.7 | 3D visualisation of the multi-wavelength sedimentation velocity data. Presented is a single time point from the 2:1 (protein:DNA) loading titration. Here, an absorbance maximum is visible below 230 nm, which is contributed by the peptide backbone of the protein, while the maximum between 260-280 nm is contributed by both the DNA and the aromatic amino acid residues of the protein.

Once analysed, the intrinsic extinction profile of the individual components can be used to decompose the multi-wavelength sedimentation velocity data into separate sedimentation profiles, which are scaled to molar concentration—this is achieved by employing the non-negatively constrained least-squares algorithm (12; 14). An example of these separated signals is presented in Figure 6.8 where a clear difference can be visualised in the sedimentation boundaries between DNA and protein. Following decomposition, the formation of a hetero-complex can be identified by observing a co-migrating peak (or signal) between the individual sedimentation distributions.

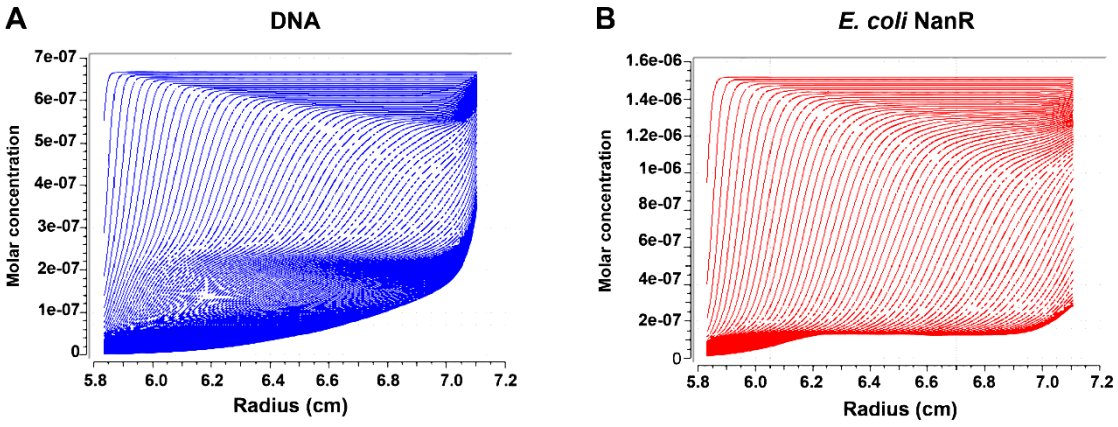


Figure 6.8 | 2D visualisation of the *E. coli* NanR and DNA sedimentation profiles following spectral decomposition. Presented is the separated signals from the 2:1 (protein:DNA) titration, where the data has been scaled to molar concentration as a function of radius. Here, a clear difference can be visualised in the sedimentation boundaries between DNA and *E. coli* NanR. While NanR is primarily one large boundary, thus a single species, DNA has multiple boundaries or several species.

Finally, the stoichiometry of the hetero-complex can simply be extracted by integrating the molar ratio of the co-migrating peaks (11). Moreover, these separated datasets can now be analysed further to extract hydrodynamic information as is traditionally achieved using single-wavelength experiments. Thus, this new approach can provide both hydrodynamic and spectral characterisation of an interacting system to define the stoichiometry of association, and additionally grant access to properties such as the mass and frictional ratio of each species—together this enables interacting species to be characterised with high resolution.

Overall, the multi-wavelength sedimentation velocity analysis presented in Chapter Five verifies the stoichiometry that was defined in the single-wavelength experiments above. This demonstrates the precision of analytical ultracentrifugation and reinforces its recognition as the gold-standard technique to quantitatively analyse complex interactions in solution (12; 15). However, most importantly it showcases what multi-wavelength sedimentation velocity analysis can achieve with the additional information that the spectral dimension can offer.

6.2.4 The C-terminal domain sub-structure of *E. coli* NanR used to phase the native structure

Chapter Five reports the crystal structure of *E. coli* NanR solved in complex with Neu5Ac and a zinc ion to a maximum resolution of 2.1 Å. To phase this structure, a combined single-wavelength anomalous dispersion and molecular replacement strategy was employed, using zinc as the anomalous scatterer. This experimental phasing approach was implemented following identification of a zinc ion through inductively coupled plasma mass spectrometry and X-ray absorption spectroscopy.

Initially when this anomalous signal for zinc was detected, the reported value was low in both the fluorescence scan and multiple anomalous dispersion scan, suggesting poor occupancy. To address this problem, NanR was expressed in 10 µM zinc chloride, purified and subjected to further crystallisation trials. This resulted in the growth of a large orthorhombic crystal that diffracted to a maximum resolution of 2.29 Å (Figure 6.9A-B). Due to the size, multiple datasets could be collected along the length of the crystal at a wavelength of 1.278 Å (9.7 keV), which is slightly off centre from the absorption edge of zinc (1.284 Å). Upon merging these datasets, a significant anomalous signal was observed for zinc. This anomalous signal was suitable for experimentally phasing the merged data using Auto-Rickshaw (16).

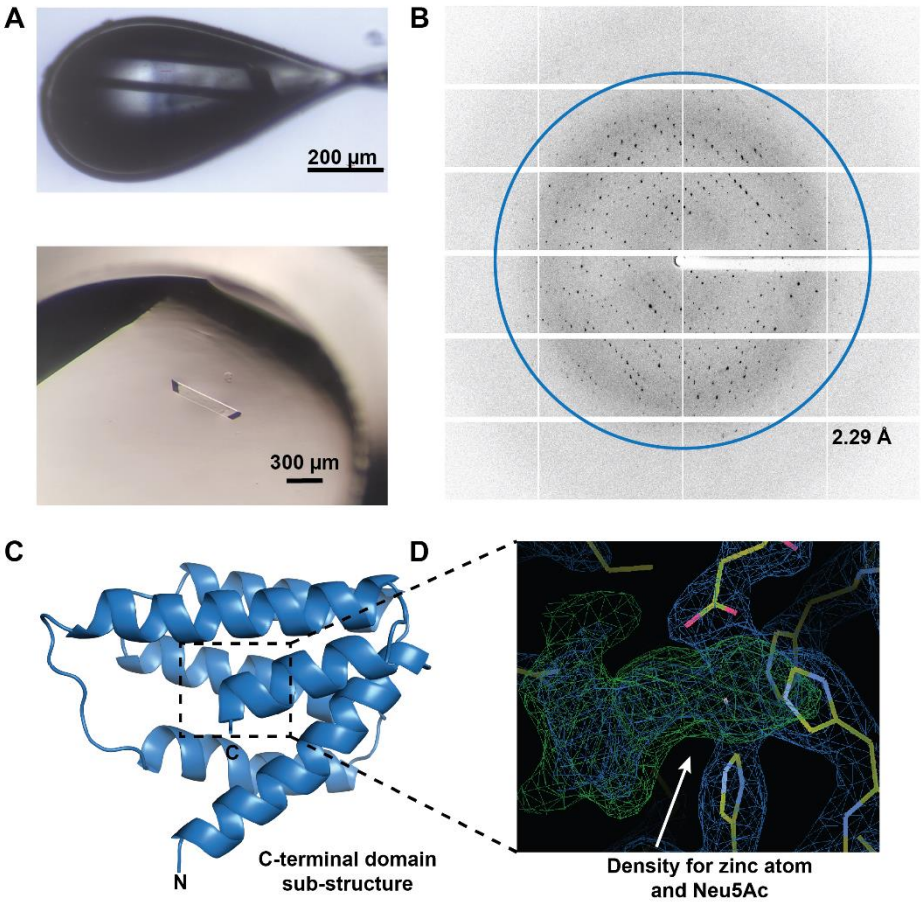


Figure 6.9 | *E. coli* NanR substructure. A) Image of the crystal in the loop (top) and within the crystallisation plate (lower). A scale bar for each image is shown as a black line. **B)** X-ray diffraction image. Blue ring highlights maximum resolution of 2.29 Å. **C)** Cartoon representation of the substructure, where the N- and C-terminals are labelled. **D)** The strong electron density for zinc and partial ligand density. Figures produced using PyMOL and *COOT*.

Following further refinement, the final model only showed density for the C-terminal domain (residues 119 to 244) (Figure 6.9C). This suggests that the N-terminal domain has been cleaved due to proteolysis, as chymotrypsin was included in the crystallisation condition. However, most importantly, strong density was observed for the zinc atom, coordinated by three histidine residues and a glutamic acid residue (Figure 6.9D). Partial density was also observed for a ligand that appears to be interacting with the zinc atom. In the native dataset this ligand was unambiguously identified as Neu5Ac. Although not a complete model, this substructure was fundamental in solving the native dataset as molecular replacement alone was not viable as a phasing strategy. This is because the C-terminal domain among the GntR family displays significant structural diversity.

6.2.5 Crystallisation trials of *E. coli* NanR in the presence of DNA

Following sedimentation velocity experiments to define the stoichiometry of the NanR-DNA hetero-complex (Chapter Five and Section 6.2.3), *E. coli* NanR was subjected to crystallisation trials in the presence of DNA with the aim of obtaining a crystal structure to elucidate the specific DNA contacts. Here, trials were conducted using the sitting-drop vapor-diffusion method across the crystallisation screens Shotgun and MIDAS*plus*, a screen optimised for macromolecular complexes (Molecular Dimensions).

6.2.5.1 Crystallisation trials with full-length NanR

In an initial strategy, full-length protein was pre-equilibrated with either two or three binding site oligonucleotides (label-free) and then subjected to co-crystallisation trials. The mixing ratios for each oligonucleotide were selected based on the appearance of a stable species in the multi-wavelength sedimentation velocity experiments (Chapter Five). Unfortunately, these trials were largely unsuccessful, despite some appearance of precipitation and phase separation, which was likely due to heterogeneity in the sample.

To address this heterogeneity, I chose to focus on the two binding site model, which is a simpler system. In addition, the protein-DNA hetero-complex was isolated from free protein and DNA, by subjecting the sample to size exclusion chromatography prior to setting up the co-crystallisation trials. This was facilitated by utilising the multi-wavelength detection on the ÄKTA pure (GE Healthcare), which enabled a few mixing ratios to be screened that were optimised for the dimeric or tetrameric NanR-DNA hetero-complexes. Furthermore, the DNA oligonucleotide was redesigned to have sticky ends, as opposed to blunt ends (Figure 6.10A). These additional base pairs are believed to encourage head-to-tail polymerisation of the DNA to form an extended chain and assist crystal packing (17). Following several weeks of growth, small crystals were obtained for NanR with this new oligonucleotide at a mixing ratio of 2:1 (protein:DNA) in Shotgun condition B11 (0.2 M Ammonium chloride, 20 % (w/v) PEG 3350). When data was collected at the MX2 beamline (Australian Synchrotron), these crystals diffracted to approximately 13 Å (Figure 6.10B-C), but the data was unable to be processed, and thus no spacegroup was obtained.

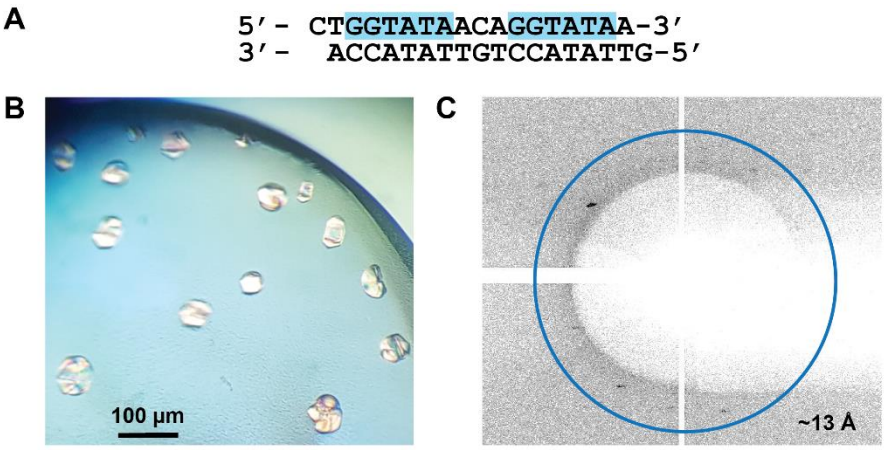


Figure 6.10 | *E. coli* NanR-DNA crystallisation trials. A) Two binding site model with 'sticky ends'. Highlighted in blue are the two GGTATA binding sites. **B)** Image of the crystals within the crystallisation drop. **C)** X-ray diffraction image. Blue ring indicates the estimated resolution of ~13 Å.

6.2.5.2 Crystallisation trials with the N-terminal DNA-binding domain

While optimisation of the crystallisation conditions is ongoing with the full-length protein and DNA, an alternative, parallel strategy was employed utilising just the N-terminal DNA-binding domain of *E. coli* NanR. Briefly, this was achieved through PCR and In-Fusion cloning of the *E. coli nanR* gene in order to produce a template that encoded only the N-terminal domain (residues 1-95) and also incorporate an N-terminal His-tag by utilising the restriction sites *Nde*I and *Xho*I in the expression vector pET28a (Chapter Eight, Section 8.6.3.6). Following In-Fusion cloning, the NanR construct was transformed into *E. coli* BL21 (DE3) bacterial cells and overexpressed in Luria Bertani media. As the construct contained an N-terminal His-tag, a two-step purification procedure was employed, which included immobilised metal affinity chromatography and size-exclusion chromatography (as described in Chapter Eight, Section 8.6.3.7). As this truncated construct does not contain any aromatic residues, protein elution during each chromatography technique was monitored *via* the peptide backbone absorbance at a wavelength of 220 nm. Alternatively, a Bradford assay was used to estimate protein concentration. The increasing purity of the truncated protein, following each chromatography technique is shown in Figure 6.11A. The final purity was estimated to be approximately 90-95%, with only a minor contaminant at ~9 kDa.

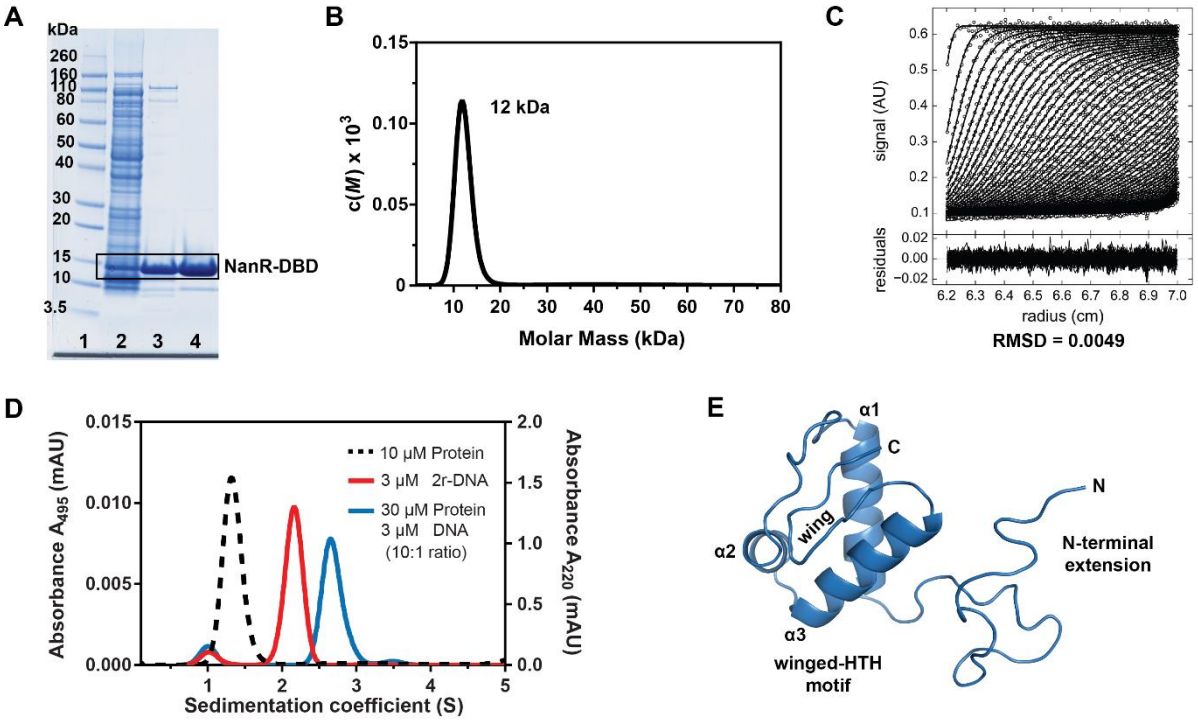


Figure 6.11 | *E. coli* NanR N-terminal DNA-binding domain truncation. A) SDS-PAGE analysis showing the increasing purity following each purification step. Lane 1, Protein ladder (kDa); lane 2, crude lysate; lane 3, pooled fractions from immobilised-metal affinity chromatography; and lane 4, pooled fractions following size-exclusion chromatography. **B)** The c(M) model displaying the molar mass distribution. Shown in bold is the apparent molar mass, following peak integration. **C)** The raw data (open circles) and best fit (lines) is presented. For clarity, every third absorbance scan is shown for each dataset. The randomly distributed residuals for each fit are shown below. **D)** 2DSA-Monte Carlo model displaying the sedimentation coefficient (S) distribution from the 55 000 rpm experiment, where each colour represents a different titration sample, collected independently (see legend). Samples with DNA were collected at 495 nm, while protein alone (black dash) was collected at 280 nM. **E)** Cartoon representation of the N-terminal DNA-binding domain (residues 1-95) based off the atomic structure in Chapter Five. Highlighted is the N-terminal extension (disordered), and the key features of the winged helix-turn-helix (HTH) motif.

To characterise the stability of the N-terminal DNA-binding domain in solution, sedimentation velocity experiments were conducted, using analytical ultracentrifugation (Chapter Eight, Section 8.6.3.4). When fitted to a continuous mass distribution [c(*M*)] model, the data gave an apparent molar mass of 12 kDa, consistent with the theoretical monomeric mass of 12.9 kDa (Figure 6.11B). Validation of this analysis was supported by the randomly distributed residuals and a low root-mean-square deviation (Figure 6.11C). Further fitting to a continuous sedimentation coefficient distribution [c(*s*)] model, displayed a single peak with a sedimentation coefficient of 1.34 S (Figure 6.11D), supporting a homogeneous species in solution. The frictional ratio (f/f_0) of 1.4 suggests this truncated construct has an elongated shape in solution. This is a feature of the disordered N-terminal extension and His-tag (Figure 6.11E).

To confirm the activity of the N-terminal DNA-binding domain, prior to crystallisation trials, further sedimentation velocity experiments were conducted using 10:1 ratio of protein to fluorescently end-labelled DNA at 495 nm (Chapter Eight, Section 8.6.3.4). Following data collection at 55 000 rpm, a positive shift in sedimentation coefficient relative to the individual interaction partners was observed (Figure 6.11D). This was consistent with hetero-complex formation, and thus provides confirmation of DNA-binding activity.

Having confirmed the N-terminal DNA-binding domain was active, crystallisation trials were conducted using the two binding site model with 'sticky ends' (Figure 6.10A) at a protein:DNA ratio of 10:1. Following several weeks growth, small crystals were observed in four different conditions (Figure 6.12). These include: A) Shotgun condition D9 (0.2 M Ammonium sulfate, 0.1 M sodium acetate, pH 4.6, and 30 % (w/v) PEG 2000 MME); B) Shotgun condition F3 (0.2 M Ammonium sulfate, 20 % (w/v) PEG 3350); C) MIDAS*plus* condition H12 (0.1 M Tris, pH 8.0, 15 % (w/v) Polyvinylpyrrolidone, and 25 % (w/v) PEG 5000 MME; and D) Shotgun condition B11 (0.2 M Ammonium chloride, and 20 % (w/v) PEG 3350). Although small, the crystals all appeared to have a similar oval morphology with more rounded edges. Overall, these crystals are suitable for optimisation. One strategy that would be useful to employ in the future is crystal seeding, with the aim of increasing the size of the crystal.

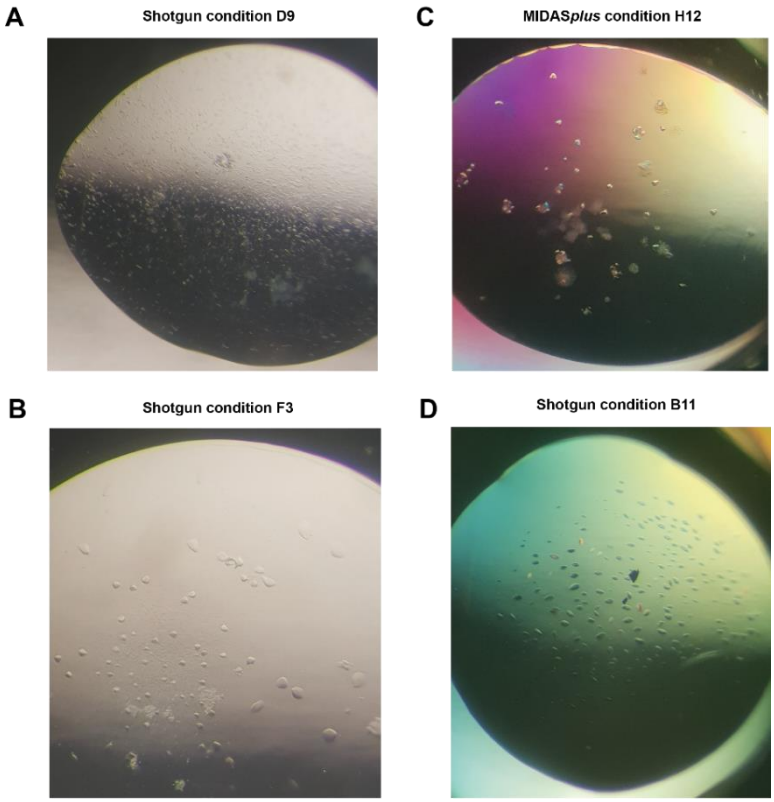


Figure 6.12 | *E. coli* NanR N-terminal DNA-binding domain-DNA crystallisation trials. A) Shotgun condition D9. **B)** Shotgun condition F3. **C)** MIDAS*plus* condition H12. **D)** Shotgun condition B11. Sitting-vapour diffusion drops were visualised using a Leica benchtop microscope. A polarising filter was used for both C and D.

6.3 Chapter Summary

This chapter provided supplementary structural and biophysical data, to both support the evidence presented in Chapter Five, and enhance our understanding of the GntR-type transcriptional regulator NanR from *E. coli*.

Nano-DSF experiments demonstrated that in the presence of Neu5Ac, an increase in thermal stability was observed. This showed that Neu5Ac can actively bind NanR in solution, supporting the hypothesis that this is the effector molecule in *E. coli*. In the presence of DNA, a significant increase in thermal stability was observed, supporting hetero-complex formation.

Although the DNA binding kinetics were explored in Chapter Five, this chapter investigated if the distance between the direct GGTATA repeats was crucial for binding, by progressively increasing the number of base pairs. EMSA experiments illustrated that mutating the distance between direct repeats two and three promote a significant decrease in DNA-binding affinity. This suggested that the third GGTATA repeat was crucial for the assembly of higher order hetero-complexes.

Single-wavelength sedimentation velocity studies with DNA demonstrated that when two GGTATA binding sites are present, a dimeric and tetrameric NanR-DNA hetero-complex was formed. In the presence of the full operator site, a tetrameric and hexameric NanR-DNA hetero-complex was formed. This was verified by an excellent agreement between the apparent molar mass of these hetero-complexes and their respective theoretical molar mass. Interestingly, the dimeric NanR-DNA hetero-complex was not definitely resolved with the full operator site, despite being observed in the two binding site experiment. This illustrated that NanR rapidly assembles into the higher order hetero-complexes, and although present in solution, the signal concentration of the dimeric NanR-DNA hetero-complex fell outside the detectable range of the instrument. Therefore, multi-wavelength sedimentation velocity experiments were carried out, as the additional spectral dimension allowed this hetero-complex to be resolved, when using the full operator site. Nonetheless, the single-wavelength experiments presented in this chapter provided the foundation to design the multi-wavelength experiment.

The C-terminal domain sub-structure that was used to phase the native dataset and solve the final model of NanR was presented. Analogous to this final model, the sub-structure also displayed density for the ligand Neu5Ac, which interacts with the zinc ion. Following preliminary crystallisation trials, crystals were obtained for NanR in the presence of DNA with ‘sticky ends’, which was used to encourage crystal packing. Unfortunately, while diffraction data was collected for these complex crystals, the maximum resolution of the diffraction was ~13 Å. In addition, various crystals were obtained with the

N-terminal DNA-binding domain and DNA, following the cloning, purification and preliminary characterisation of this domain construct. Together, these crystals should be optimised in the future.

Overall, these data complement the characterisation presented in Chapter Five, enhance our understanding of the GntR-type NanR from *E. coli*, and provide a platform for future structure determination of the protein-DNA hetero-complex at an atomic level. This would allow the specific amino acid residues that interact with DNA to be deduced and may additionally provide further insight into the mechanistic role of the N-terminal extension in mediating the cooperative binding to DNA.

6.4 Chapter References

1. Seabrook, S. A., & Newman, J. (2013). High-Throughput Thermal Scanning for Protein Stability: Making a Good Technique More Robust. *ACS Combinatorial Science, 15*, 387-392.
2. Greenfield, N. J. (2006). Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols, 1*, 2876-2890.
3. Burstein, E. A., Vedenkina, N. S., & Ivkova, M. N. (1973). Fluorescence and the location of tryptophan residues in protein molecules. *Photochemistry and Photobiology, 18*, 263-279.
4. Haffke, M., Rummel, G., Boivineau, J., Münch, A., & Jaakola, V.-P. (2016). *nanoDSF: Label-free Thermal Unfolding Assay of G Protein-Coupled Receptors for Compound Screening and Buffer Composition Optimization*.
5. Martin, L., Schwarz, S., & Breitsprecher, D. (2015). *Prometheus: the platform for analyzing protein stability and thermal unfolding of proteins*.
6. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
7. Brookes, E., Cao, W. M., & Demeler, B. (2010). A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *European Biophysics Journal with Biophysics Letters, 39*, 405-414.
8. Demeler, B. (2005). UltraScan - A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments. In D. J. Scott, S. E. Harding, & A. J. Rowe (Eds.), *Analytical Ultracentrifugation: Techniques and Methods* (pp. 210-230): The Royal Society of Chemistry.
9. Brookes, E. H., & Demeler, B. (2007). *Parsimonious regularization using genetic algorithms applied to the analysis of analytical ultracentrifugation experiments*. Paper presented at the Proceedings of the 9th annual conference on Genetic and evolutionary computation, London, England.
10. Pearson, J. Z., Krause, F., Haffke, D., Demeler, B., Schilling, K., & Colfen, H. (2015). Next-Generation AUC Adds a Spectral Dimension: Development of Multiwavelength Detectors for the Analytical Ultracentrifuge. *Methods in Enzymology, 562*, 1-26.
11. Gorbet, G. E., Pearson, J. Z., Demeler, A. K., Colfen, H., & Demeler, B. (2015). Next-Generation AUC: Analysis of Multiwavelength Analytical Ultracentrifugation Data. *Methods in Enzymology, 562*, 27-47.
12. Gorbet, G. E., Pearson, J. Z., Demeler, A. K., Colfen, H., & Demeler, B. (2015). Next-Generation AUC: Analysis of Multiwavelength Analytical Ultracentrifugation Data. In J. L. Cole (Ed.), *Analytical Ultracentrifugation* (Vol. 562, pp. 27-47).
13. Schuck, P., Perugini, M. A., Gonzales, N. R., Howlett, G. J., & Schubert, D. (2002). Size-Distribution Analysis of Proteins by Analytical Ultracentrifugation: Strategies and Application to Model Systems. *Biophysical Journal, 82*, 1096-1111.
14. Lawson, C. L., & Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs, NJ., Prentice-Hall, Inc.
15. Lebowitz, J., Lewis, M. S., & Schuck, P. (2002). Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein science : a publication of the Protein Society, 11*, 2067-2079.
16. Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S., & Tucker, P. A. (2005). Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. *Acta Crystallographica Section D, 61*, 449-457.
17. Russo Krauss, I., Merlino, A., Vergara, A., & Sica, F. (2013). An overview of biological macromolecule crystallization. *International Journal of Molecular Sciences, 14*, 11643-11691.

Chapter Seven

Conclusion and future perspectives

7.1 The interplay between gene regulation and metabolism in pathogenic bacteria

Within the human host, the variety of environments that pathogenic bacteria infect (e.g. gastrointestinal or respiratory tract) present different nutrient availability (1; 2). To gain a competitive advantage, pathogens must adapt to these unfavourable environments by upregulating the expression of genes that are essential to metabolise the available nutrient source. This response is facilitated by transcriptional regulation. Here, in the absence of said nutrient, expression of the associated metabolic machinery is switched off by the binding of a transcriptional regulator to the operator region of the metabolic operon (3)—this blocks the activity of RNA polymerase. Although when sufficient levels of said nutrient are present, signalling molecules called effectors can bind to the transcriptional regulator and mediate a conformational change through an allosteric mechanism (4; 5)—as a result of this environmental cue, gene expression is switched on. Together, this complex interplay between gene expression and the metabolic network is critical for efficient growth and to establish infection within the human host. Furthermore, as this coordination provides pathogenic bacteria the ability to propagate and persist in new niches, these nutrient sources can act as virulence factors and thus play an integral role in pathogenesis (6; 7).

Sialic acid is an example of such nutrient that acts as a virulence factor in pathogenic bacteria (8-10). Found at the terminal of glycoconjugates within the human respiratory and gastrointestinal tract (10; 11), the utilisation of sialic acid provides bacteria with an alternative source of carbon, nitrogen and energy (9; 12). The expression of the molecular machinery to catabolise sialic acid is controlled by the transcriptional regulator, NanR. However, the mechanism of this gene regulation is poorly understood at a molecular level. Using protein biophysics and structural biology, I addressed this gap by investigating two types of NanR sialoregulators, including the RpiR-type sialoregulator from the Gram-positive pathogen *S. pneumoniae* and the GntR-type sialoregulator from the Gram-negative pathogen *E. coli*. In addition, I also investigated the uncharacterised *yjhBC* operon, which is hypothesised to play a role in sialic acid metabolism. Together, this knowledge provides a foundation to dissect the regulatory mechanism of NanR at a molecular level, and most importantly enhance our understanding of sialic acid catabolism to assist in the development of novel antimicrobial therapeutics that aim to prevent bacterial colonisation and persistence.

7.2 The first molecular insight into the *yjhBC* operon

Within bacteria, the catabolism of host-derived sialic acid primarily refers to Neu5Ac, the most ubiquitous form of sialic acid in the human host (9; 13). Aside from Neu5Ac, the 9-O-acetylated derivative and the hydroxylated derivative, *N*-glycolylneuraminic acid are reported as nutrient sources. These sialic acid derivatives are processed by the periplasmic esterase NanS (14; 15) and the epimerase YhcH (16), respectively. In addition, as part of the NanR coordinated regulation in *E. coli*, a putative oxidoreductase YjhC and a putative permease YjhB have since been identified, which are hypothesised to uptake and process less common forms of sialic acid (9). Prior to this thesis, structural and functional data to support this involvement in sialic acid catabolism was lacking.

Excitingly, I presented the first molecular insight into the previously uncharacterised *yjhBC* operon. The high-resolution structure of *E. coli* YjhC to 1.35 Å, solved in complex with NAD(H) verifies its role as an oxidoreductase/dehydrogenase from the Gfo/Idh/MocA family. Despite a low sequence identity, YjhC shares a similar tertiary fold to family members that catalyse redox reactions with carbohydrate-based substrates. This annotation was supported through *in vivo* knockout studies. Thermal shift experiments and *in silico* docking simulations, identified several promising leads, including the sialic acid 1,7 lactone derivative and Neu5Ac. In terms of relative distribution, the lactone derivative is the second most abundant sialic acid in the human gastrointestinal tract with respect to Neu5Ac (9; 17). Unfortunately, soaking experiments failed to observe any density for either ligand when using the NAD(H) co-crystals. In the native structure a mysterious density was observed that was localised around the proposed catalytic site. This feature mostly likely represents a bound substrate and provides a possible explanation as to why ligand density was not observed following soaking experiments. Nonetheless, these results are consistent with an oxidoreductase/dehydrogenase functionality for YjhC and provide evidence to suggest YjhC is involved in sialic acid catabolism, directed towards Neu5Ac derivatives. Importantly, the derivatives identified as promising leads have abundance levels in the human host that are comparable to Neu5Ac. If these levels were considerably different, it would question the fitness benefit towards coordinating the gene regulation of these enzymes in *E. coli*.

I also investigated the permease YjhB, and optimised an overexpression protocol, using a protein-GFP fusion system. This is a critical step towards purifying the membrane protein for future structural and biophysical studies. A *de novo* 3D reconstruction and subsequent *in silico* characterisation suggested YjhC is a sugar-ion symporter and identified amino acid residues that may be of functional importance. Interestingly, in contrast with other sugar-ion symporters, such as the fructose transporter (18) or the *Proteus mirabilis* SiaT (19), YjhB does not contain a charged amino acid residue at the substrate binding site. This observation suggests YjhB does not transport a charged substrate, such as Neu5Ac. Given that

yjhB is transcribed from the same operon as *yjhC*, perhaps this symporter facilities the transport of the lactone derivative, as NanT in *E. coli* is reported to transport Neu5Ac (20; 21). An *in vivo* growth assay to test this test substrate specificity, as well as molecular dynamic simulations to further validate the model and assess the involvement of amino acid residues in the function of YjhB, would be an interesting avenue to explore for future research.

7.3 Structural and functional insight into the RpiR-type sialoregulator from *S. pneumoniae*

While RpiR-type NanR have been investigated across several bacterial pathogens including: *C. perfringens* (22); *H. influenzae* (23); *S. aureus* (24); *S. pneumoniae* (25; 26); and *V. vulnificus* (27; 28); the focus of this research was primarily aimed at determining the DNA sequence that is recognised by NanR and identifying which molecules can induce gene expression *in vivo*. As such, this research does not address the molecular mechanism of how RpiR-type sialoregulators control gene expression, except for NanR from *V. vulnificus*.

To address this, I investigated the RpiR-type NanR from *S. pneumoniae* using protein biophysics and structural biology. Through sedimentation velocity experiments and fluorescently end-labelled DNA that replicated the recognition site, a 2:1 binding stoichiometry and the first dissociation constant for *S. pneumoniae* NanR were reported. This DNA-binding affinity was consistent with the reported dissociation constant for NanR from *V. vulnificus* (28), which was determined using isothermal titration calorimetry. While both these techniques can analyse the interaction in solution, isothermal titration calorimetry is a laborious method and requires a considerably larger amount of sample than sedimentation velocity experiments. Furthermore, the signal in isothermal titration calorimetry is proportional to the binding enthalpy of the interaction, thus if the protein-DNA interaction exhibits a small-binding enthalpy this will result in low signal-to-noise (29). Conversely, sedimentation velocity experiments using an analytical ultracentrifuge present a more convenient method, which can provide access to both thermodynamic and hydrodynamic properties, simply by analysing the sedimentation and diffusion behaviour of macromolecules in solution (30).

Small angle X-ray scattering demonstrated that *S. pneumoniae* NanR adopts an extended dimeric architecture in solution, mediated by a flexible linker region, complementing sedimentation velocity experiments. This conformation changed when forming the protein-DNA hetero-complex. Using multiphase modelling, an *ab initio* model was reconstructed for the NanR-DNA hetero-complex from the individual scattering profiles. Despite its low resolution, this model demonstrated the relative orientation of the interacting partners, placing DNA between the N-terminal DNA-binding domains of

the protein. Together, these solution studies were used to investigate the ambiguous oligomeric assembly of *V. vulnificus* NanR by comparing the experimental data to the theoretical scattering profile, which was generated from the published atomic structure (PDB ID – 4IVN). Interestingly, this comparison supported that *S. pneumoniae* NanR adopts a similar dimeric assembly in solution to the crystal structure of *V. vulnificus* NanR, when in the absence of DNA. However, the poor fit of the proposed 'active form' of *V. vulnificus* NanR, to the scattering profile of the *S. pneumoniae* NanR-DNA hetero-complex, suggested that the model of *V. vulnificus* NanR is incorrect.

Lastly, using thermal shift experiments, various effector molecules were screened to investigate binding *in vitro*. These experiments identified the sialic acid pathway metabolite *N*-acetylmannosamine-6-phosphate, which was consistent with previous *in vivo* analysis of *S. pneumoniae* NanR (25; 26). The consequence of this binding event was investigated in the presence of DNA. Interestingly, this resulted in a complete attenuation of DNA binding, strongly supporting that *N*-acetylmannosamine-6-phosphate is the effector molecule. Further structural analysis is necessary to verify this conclusion and allow any effector-induced conformational changes to be mapped. While attempts to solve the crystal structure of *S. pneumoniae* NanR were unsuccessful, diffraction data was collected between 3 Å and 4 Å. Those crystallisation conditions are perfectly suited for optimisation and should be explored in the future. Nonetheless, this characterisation permitted a model mechanism for *S. pneumoniae* NanR-mediated gene regulation to be proposed.

7.4 The cooperative assembly of the GntR-type NanR from *E. coli*

I also conducted a thorough structural and functional investigation of the GntR-type NanR from *E. coli*. In contrast to the RpiR-type sialoregulators, which function as a simple system, the GntR-type NanR from *E. coli* is significantly more complex. Using novel multi-wavelength sedimentation velocity analysis, I demonstrated that *E. coli* NanR binds DNA as a dimer to a total of three direct GGTATA repeats, forming a hexameric NanR-DNA hetero-complex, with a 6:1 stoichiometry. This is the first application of this emerging technique with protein and DNA as interaction partners in solution. Further binding analysis demonstrated this assembly process was cooperative and reported a nanomolar dissociation constant. Although cooperativity among bacterial transcriptional regulators has long been established, in such examples as the well-known *lac* repressor (31) or *ara* repressor (32), this behaviour has not been reported for any other NanR gene regulator of sialic acid catabolism. The mechanisms that drive cooperativity are diverse, facilitated by protein-protein interaction, the presence of DNA or through indirect processes that occur locally or over a considerable distance (33). To address the driving force of this cooperativity in *E. coli* NanR, I initially conducted a sequence analysis. This identified a 32 residue

N-terminal extension, which is considerably longer than closely related structures of GntR transcriptional regulators and therefore unique to *E. coli*. Using mutagenesis, the involvement of this N-terminal extension in DNA binding was assessed. Curiously, these experiments revealed that the ability to form higher-order hetero-complexes was perturbed when using this truncated construct of NanR. This provided strong evidence that the N-terminal extension mediates the positive cooperativity of the *E. coli* sialoregulator, a process that is likely facilitated by protein-protein interactions.

The DNA recognition site for *E. coli* NanR is unique among all characterised members of the GntR superfamily. Although some members of the FadR and HutC subfamilies are reported to contain inverted or direct repeats, these recognition sites contain a maximum of two repeats (4; 34), whereas *E. coli* NanR contains three exact (or near-exact) direct repeats of the hexanucleotide sequence GGTATA (35). Combined with the positive cooperativity observed for this sialoregulator, perhaps this feature is the result of evolution in order to maintain robust control over the gene expression of sialic acid catabolic machinery. This is because the binding to a single GGTATA repeat exhibited poor DNA-binding affinity. However, the presence of multiple GGTATA sites demonstrated a dynamic hetero-complex assembly process. To investigate the importance of the distance between the direct repeats, I designed mutant oligonucleotides and conducted EMSA experiments. Interestingly, this data showed that the pairs of direct GGTATA repeats are not equivalent and suggested that cooperativity of the system is mediated by direct GGTATA repeats one and two. However, it is unclear if the binding of NanR occurs sequentially to these repeats or through a random-ordered process. This is a question that should be explored in future experiments.

Excitingly, this thesis reported the crystal structure of *E. coli* NanR to 2.1 Å, solved in complex with Neu5Ac and a zinc ion. This coordination supports the role of Neu5Ac as the effector molecule, while additionally providing the first report of a metal ion playing a direct role in the binding of an effector molecule, within the GntR superfamily. Mutated variants would allow the residues that were involved in the coordination of Neu5Ac and zinc to be probed, which I plan to do in the future. Using single-particle cryo-EM, the structure of the dimeric NanR-DNA hetero-complex was refined to a resolution of 3.9 Å, providing the first structural insight into the protein-DNA conformation for a NanR gene regulation. In addition, this conformation illustrated the significant rearrangement of the N-terminal DNA-binding domains, which was facilitated by the flexible α4 linker of NanR. Remarkably, given the molar mass of the complex (70.5 kDa), this experiment did not employ a phase plate for data collection. A search through the Electron Microscopy Data Bank (EMDB) identified this protein-DNA hetero-complex as the highest resolution model to data, given its size, lack of symmetry and dynamic nature. All the high-resolution structures below 100 kDa are rigid, homogeneous proteins, such as: haemoglobin (54 kDa, C2 symmetry reconstruction, resolved to 3.2 Å) (36); and isocitrate dehydrogenase (93 kDa, C2

symmetry, resolved to 3.8 Å) (37). Further cryo-EM experiments of the higher order hetero-complex provided structural evidence to support the multimeric assembly process and highlighted how close each NanR dimer are to each other, when bound to DNA. This proximity reinforces the notion that protein-protein interactions are the driving force of cooperativity, where perhaps the N-terminal extension helps to tether the sequential dimers together. While DNA-mediated interactions cannot be ruled out, no apparent distortion in the DNA was visible, nor any suggestion of DNA looping. Higher resolution cryo-EM models of the higher order hetero-complexes may provide further insight towards how the N-terminal extension is interacting with the protein and DNA.

Although this detailed characterisation permits a model mechanism for regulation to be proposed, one unresolved question remains—what are the molecular determinants that facilitate the allosteric regulation of NanR? I identified zinc as an additional effector molecule and surprisingly captured the structure of NanR in an effector-bound and effector-free (apo) state. Superimposition of these states provided insight into the molecular choreography that occurs upon binding of Neu5Ac and zinc. However, only a small attenuation of DNA-binding activity can be observed *in vitro* with these molecules. This insignificant change shows that the allosteric mechanism involves additional elements that have yet to be elucidated *in vitro*. One hypothesis is DNA methylation, which although presents broad functionality in bacteria, the activity of DNA adenine methyltransferase (Dam) has been reported to influence the transcription of virulence factors and thus play a role in pathogenesis (38). Located immediately upstream of all three regulated operons in *E. coli* is the sequence GATC, which serves as a consensus site for Dam (10; 39). Interestingly, one study reported a decrease in expression of sialic acid catabolism in a Dam mutant (40). By consequence, it is hypothesised that methylation over time may reduce the ability for NanR to rebind in the presence of Neu5Ac (10; 39), although this has yet to be confirmed experimentally. An alternative hypothesis is an interplay between the cAMP receptor protein (CRP), which facilitates the binding of RNA polymerase to the promoter region of DNA through protein-protein interaction (41). In *E. coli*, ~60 bp upstream of the *nanATEK* operon there is a CRP binding site (42), which is commonly observed in carbon catabolite operons (41).

7.5 Chapter References

1. Jeong, H. G., Oh, M. H., Kim, B. S., Lee, M. Y., Han, H. J., & Choi, S. H. (2009). The capability of catabolic utilisation of *N*-acetylneuraminic acid, a sialic acid, is essential for *Vibrio vulnificus* pathogenesis. *Infection and Immunity, 77*, 3209-3217.
2. Chang, D. E., Smalley, D. J., Tucker, D. L., Leatham, M. P., Norris, W. E., Stevenson, S. J., Anderson, A. B., Grissom, J. E., Laux, D. C., Cohen, P. S., & Conway, T. (2004). Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 7427-7432.
3. Desvergne, B., Michalik, L., & Wahli, W. (2006). Transcriptional regulation of metabolism. *Physiological Reviews, 86*, 465-514.

4. Jain, D. (2015). Allosteric control of transcription in GntR family of transcription regulators: A structural overview. *IUBMB Life, 67*, 556-563.
5. Ptashne, M. (2004). *A genetic switch : phage lambda revisited* (3rd ed.). Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
6. Rohmer, L., Hocquet, D., & Miller, S. I. (2011). Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiology, 19*, 341-348.
7. Schmidt, H., & Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clinical and Microbiology Reviews, 17*, 14-56.
8. Bouchet, V., Hood, D. W., Li, J., Brisson, J. R., Randle, G. A., Martin, A., Li, Z., Goldstein, R., Schweda, E. K., Pelton, S. I., Richards, J. C., & Moxon, E. R. (2003). Host-derived sialic acid is incorporated into *Haemophilus influenzae* lipopolysaccharide and is a major virulence factor in experimental otitis media. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 8898-8903.
9. Vimr, E. R. (2013). Unified theory of bacterial sialometabolism: how and why bacteria metabolise host sialic acids. *International Scholarly Research Notices Microbiology, 2013*, 816713.
10. Vimr, E. R., Kalivoda, K. A., Deszo, E. L., & Steenbergen, S. M. (2004). Diversity of microbial sialic acid metabolism. *Microbiology and Molecular Biology Reviews, 68*, 132-153.
11. Baos, S. C., Phillips, D. B., Wildling, L., McMaster, T. J., & Berry, M. (2012). Distribution of sialic acids on mucins and gels: A defense mechanism. *Biophysical Journal, 102*, 176-184.
12. Almagro-Moreno, S., & Boyd, E. F. (2009). Insights into the evolution of sialic acid catabolism among bacteria. *BioMed Central Evolutionary Biology, 9*, 118.
13. Severi, E., Hood, D. W., & Thomas, G. H. (2007). Sialic acid utilisation by bacterial pathogens. *Microbiology, 153*, 2817-2822.
14. Rangarajan, E. S., Ruane, K. M., Proteau, A., Schrag, J. D., Valladares, R., Gonzalez, C. F., Gilbert, M., Yakunin, A. F., & Cygler, M. (2011). Structural and enzymatic characterization of NanS (YjhS), a 9-O-Acetyl N-acetylneuraminic acid esterase from *Escherichia coli* O157:H7. *Protein Science, 20*, 1208-1219.
15. Steenbergen, S. M., Jirik, J. L., & Vimr, E. R. (2009). YjhS (NanS) Is Required for Escherichia coli To Grow on 9-O-Acetylated N-Acetylneuraminic Acid. *Journal of Bacteriology, 191*, 7134-7139.
16. Teplyakov, A., Obmolova, G., Toedt, J., Galperin, M. Y., & Gilliland, G. L. (2005). Crystal Structure of the Bacterial YhcH Protein Indicates a Role in Sialic Acid Catabolism. *Journal of Bacteriology, 187*, 5520-5527.
17. Robbe, C., Capon, C., Maes, E., Rousset, M., Zweibaum, A., Zanetta, J. P., & Michalski, J. C. (2003). Evidence of regio-specific glycosylation in human intestinal mucins: presence of an acidic gradient along the intestinal tract. *Journal of Biological Chemistry, 278*, 46337-46348.
18. Dang, S., Sun, L., Huang, Y., Lu, F., Liu, Y., Gong, H., Wang, J., & Yan, N. (2010). Structure of a fucose transporter in an outward-open conformation. *Nature, 467*, 734-738.
19. Wahlgren, W. Y., Dunevall, E., North, R. A., Paz, A., Scalise, M., Bisignano, P., Bengtsson-Palme, J., Goyal, P., Claesson, E., Caing-Carlsson, R., Andersson, R., Beis, K., Nilsson, U. J., Farewell, A., Pochini, L., Indiveri, C., Grabe, M., Dobson, R. C. J., Abramson, J., Ramaswamy, S., & Friemann, R. (2018). Substrate-bound outward-open structure of a Na(+)-coupled sialic acid symporter reveals a new Na(+) site. *Nature Communications, 9*, 1753.
20. North, R. A., Horne, C. R., Davies, J. S., Remus, D. M., Muscroft-Taylor, A. C., Goyal, P., Wahlgren, W. Y., Ramaswamy, S., Friemann, R., & Dobson, R. C. J. (2018). "Just a spoonful of sugar...": import of sialic acid across bacterial cell membranes. *Biophysical Reviews, 10*, 219-227.
21. Vimr, E. R., & Troy, F. A. (1985). Identification of an inducible catabolic system for sialic acids (nan) in *Escherichia coli*. *Journal of Bacteriology, 164*, 845-853.
22. Therit, B., Cheung, J. K., Rood, J. I., & Melville, S. B. (2015). NanR, a Transcriptional Regulator That Binds to the Promoters of Genes Involved in Sialic Acid Metabolism in the Anaerobic Pathogen Clostridium perfringens. *Public Library of Science One, 10*, e0133217.
23. Johnston, J. W., Zaleski, A., Allen, S., Mootz, J. M., Armbruster, D., Gibson, B. W., Apicella, M. A., & Munson, R. S., Jr. (2007). Regulation of sialic acid transport and catabolism in *Haemophilus influenzae*. *Molecular Microbiology, 66*, 26-39.
24. Olson, M. E., King, J. M., Yahr, T. L., & Horswill, A. R. (2013). Sialic acid catabolism in *Staphylococcus aureus*. *Journal of Bacteriology, 195*, 1779-1788.
25. Afzal, M., Shafeeq, S., Ahmed, H., & Kuipers, O. P. (2015). Sialic acid-mediated gene expression in *Streptococcus pneumoniae* and role of NanR as a transcriptional activator of the *nan* gene cluster. *Applied and Environmental Microbiology, 81*, 3121-3131.

26. Gualdi, L., Hayre, J. K., Gerlini, A., Bidossi, A., Colomba, L., Trappetti, C., Pozzi, G., Docquier, J. D., Andrew, P., Ricci, S., & Oggioni, M. R. (2012). Regulation of neuraminidase expression in *Streptococcus pneumoniae*. *BioMed Central Microbiology, 12*, 200.
27. Kim, B. S., Hwang, J., Kim, M. H., & Choi, S. H. (2011). Cooperative regulation of the *Vibrio vulnificus* nan gene cluster by NanR protein, cAMP receptor protein, and *N*-acetylmannosamine 6-phosphate. *Journal of Biological Chemistry, 286*, 40889-40899.
28. Hwang, J., Kim, B. S., Jang, S. Y., Lim, J. G., You, D. J., Jung, H. S., Oh, T. K., Lee, J. O., Choi, S. H., & Kim, M. H. (2013). Structural insights into the regulation of sialic acid catabolism by the *Vibrio vulnificus* transcriptional repressor NanR. *Proceedings of the National Academy of Sciences of the United States of America, 110*, E2829-2837.
29. Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., & Liu, S. Q. (2016). Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *International Journal of Molecular Sciences, 17*.
30. Cole, J. L., Lary, J. W., T, P. M., & Laue, T. M. (2008). Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. *Methods Cell Biology, 84*, 143-179.
31. Brenowitz, M., Mandal, N., Pickar, A., Jamison, E., & Adhya, S. (1991). DNA-binding properties of a lac repressor mutant incapable of forming tetramers. *Journal of Biological Chemistry, 266*, 1281-1288.
32. Mota, L. J., Sarmento, L. M., & de Sa-Nogueira, I. (2001). Control of the arabinose regulon in *Bacillus subtilis* by AraR *in vivo*: crucial roles of operators, cooperativity, and DNA looping. *Journal of Bacteriology, 183*, 4190-4201.
33. Morgunova, E., & Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology, 47*, 1-8.
34. Suvorova, I. A., Korostelev, Y. D., & Gelfand, M. S. (2015). GntR Family of Bacterial Transcription Factors and Their DNA Binding Motifs: Structure, Positioning and Co-Evolution. *Public Library of Science One, 10*, e0132618-e0132618.
35. Kalivoda, K. A., Steenbergen, S. M., & Vimr, E. R. (2013). Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *Journal of Bacteriology, 195*, 4689-4701.
36. Khoshouei, M., Radjainia, M., Baumeister, W., & Danev, R. (2017). Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat Commun, 8*, 16099.
37. Vonck, J., & Mills, D. J. (2017). Advances in high-resolution cryo-EM of oligomeric enzymes. *Current Opinion in Structural Biology, 46*, 48-54.
38. Heusipp, G., Falker, S., & Schmidt, M. A. (2007). DNA adenine methylation and bacterial pathogenesis. *International Journal of Medical Microbiology, 297*, 1-7.
39. Westphal, L. L., Sauvey, P., Champion, M. M., Ehrenreich, I. M., & Finkel, S. E. (2016). Genomewide Dam Methylation in *Escherichia coli* during Long-Term Stationary Phase. *mSystems, 1*.
40. Oshima, T., Wada, C., Kawagoe, Y., Ara, T., Maeda, M., Masuda, Y., Hiraga, S., & Mori, H. (2002). Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in *Escherichia coli*. *Molecular Microbiology, 45*, 673-695.
41. Busby, S., & Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology, 293*, 199-213.
42. Kalivoda, K. A., Steenbergen, S. M., Vimr, E. R., & Plumbridge, J. (2003). Regulation of sialic acid catabolism by the DNA binding protein NanR in *Escherichia coli*. *Journal of Bacteriology, 185*, 4806-4815.

Chapter Eight

Methods and Materials

8.1 Experimental consumables

8.1.1 Chemical consumables

Unless otherwise stated, general chemicals, including antibiotics were purchased from Sigma-Aldrich and AppliChem. Chemicals used to prepare growth media, including Luria Bertani broth base were purchased from Invitrogen, while agar was obtained from Oxoid. Isopropyl β-D-1-thiogalactopyranoside (IPTG) was sourced from Bioline. Sialic acid and all other *N*-acetyl derivatives were purchased from Carbosynth. cOmplete™ mini protease inhibitor tablets were purchased from Roche.

8.1.2 Biological consumables

Plasmid constructs harbouring genes of interest were purchased from GenScript or GenArt (Thermo Fisher Scientific). DNA oligonucleotides, primers and any associated chemical modification were purchased from Integrated DNA Technologies (IDT). All restriction enzymes were obtained from New England Biolabs. Agarose gel DNA extraction kits were supplied by Roche. DNA-Spin™ Plasmid DNA Purification Kits (miniprep) were obtained from iNtRon Biotechnology. T4 DNA Ligase and In-Fusion® HD cloning kits, including chemically competent Stellar cells™ were purchased from TaKaRa Bio USA. Bradford dye reagent was obtained from BioRad. DNA HyperLadder marker was supplied by Bioline. Novex® Sharp Pre-stained Protein Ladder, Novex® Tris-Glycine SDS sample buffer, Benchmark™ Protein Ladder and BOLT™ running buffer were purchased from Invitrogen. SYBR® Safe DNA gel strain and SYPRO® Orange dye were also obtained from Invitrogen. IF-0a inoculating fluid, tetrazolium dye mix and carbon utilisation plates (PM1 and PM2A) were all purchased from BiOLOG Inc. (Harvard, CA, USA).

8.1.3 General consumables

Precast gels required for sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and electrophoretic mobility shift assays (EMSA) were purchased from Thermo Fisher Scientific. All prepacked chromatography columns, including molecular weight standards were purchased from GE Healthcare. Syringe filters (0.22 and 0.45 µm), including eppendorf sized spin concentrators (various

molecular weight cut-off) were supplied by Merck Millipore. Large volume spin concentrators with varying molecular weight cut-off were purchased from Sartorius. Crystallisation screens, sitting-drop crystallisation plates, UV-transmissible plate seals and cryo-protection solutions were purchased from Molecular Dimensions. Cryo loops were obtained from Hampton Research. 24 deep-well growth blocks were purchased from Thermo Fisher. Breathable sealing film was purchased from Axygen. 96 well microplates and optical adhesive film were purchased from Applied Biosystems. QUANTAFOIL® R1.2/1.3 carbon support films and UltrAuFoil® R 1.2/1.3 gold support films were purchased from Quantifoil Micro Tools (Germany). Glass capillaries used for nano-differential scanning fluorimetry and microscale thermophoresis were obtained from NanoTemper.

8.2 General Methodology

8.2.1 Mass spectrometry

To accurately analyse the molecular masses of purified proteins, mass spectrometry was carried out using a Bruker MaXis 3G ultra high-resolution time of flight mass spectrometer, equipped with an electrospray ionisation source. Proteins to be analysed were diluted to approximately 1 mg mL^{-1} in MilliQ water or Tris-buffered saline. Analysis was performed by Dr. Marie Squire of the Chemistry Department at the University of Canterbury.

8.2.2 DNA sequencing

Newly cloned pET-based constructs were sequenced using next generation sequencing by Macrogen Inc., South Korea. Purified plasmids (>50 ng μL^{-1}) were supplied and used directly in the sequencing reaction, along with the appropriate primers for the forward and complementary strands, respectively. Sequencing results were aligned with the gene sequence using the software *GENEIOUS* (Biomatters, https://www.geneious.com).

8.2.3 pH measurement

The pH of buffers and other solutions was measured using an Ultrabasic Benchtop pH meter (Denver Instrument). Initially the meter was calibrated with appropriate pH standards surrounding the target pH value. The pH of solutions was then adjusted drop-wise using 1 M or 10 M HCl (or NaOH), as appropriate.

8.2.4 Centrifugation

All centrifugation was performed using a Sorvall LYNX 4000 Superspeed (Thermo Fisher), Sorvall RC 6 Plus (Thermo Fisher) or an Eppendorf Centrifuge 5430 R (MediRay).

8.2.5 Autoclave

All media and glassware for bacterial growth were autoclaved in-house for 15 min at 121 °C.

8.2.6 Visualisation of macromolecular structures

All interactive visualisation and analysis of macromolecule structures, including proteins, DNA and *ab initio* models were performed using PyMOL (1) or UCSF Chimera (2).

8.3 Microbiology

8.3.1 Bacterial strains

Four bacterial strains of *E. coli* were used throughout this research, as described in Table 8.1.

Table 8.1 | Bacterial strains used throughout this thesis. The strain name and associated genotype is presented.

E. coli strain	Application	Cloning vector
DH5α	Cloning	F⁻ *endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG* Φ80d*lacZ*ΔM15 Δ(*lacZYA-argF*) U169, *hsdR17*(r_K^- m_K^+), λ–
Stellar™	In-Fusion Cloning	F⁻ *endA1 supE44, thi-1, recA1, relA1, gyrA96, phoA*, Φ80d*lac*ZΔM15 Δ(*lac*ZYA-*argF*) U169, Δ(mrr -*hsdRMS -mcrBC*), Δ*mcrA*, λ–
BL21 (DE3)	Expression	*fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5*
Lemo21 (DE3)	Expression	*fhuA2 [lon] ompT gal (λ DE3) [dcm] hsdS/pLemo*(CamR) *λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5* pLemo = pACYC184-*PrhaBAD-lysY*

8.3.2 Constructs used in this thesis

A list of constructs used throughout this thesis are listed in Table 8.2.

Table 8.2 | Constructs used routinely throughout this thesis. The protein, species of origin, cloning vector, extinction coefficient, monomeric molecular weight and fusion tag (if applicable) are presented.

Protein	Species of origin	Cloning vector	Extinction coefficient (M^{-1} cm^{-1})	Molecular weight (Da)	Fusion tag
*Ec*YjhC	*E. coli*	pET30ΔSE*	37 930	43 048	N-term His-tag
*Ec*YjhB	*E. coli*	pWarf(-)†	70 820	44 103 (YjhB) 73 400 (Fusion)	N-term His-tag/GFP
*Sp*NanR	*S. pneumoniae*	pET30ΔSE	27 390	32 671	-
*Ec*NanR	*E. coli*	pET28a	13 980	29 524	-
*Ec*NanR (33-263)	*E. coli*	pET28a	13 980	25 625	-
*Ec*NanR-DBD (1-95)	*E. coli*	pET28a	N/D‡	12 863	N-term His-tag

* Kindly supplied by Dr. Hironori Suzuki
† Kindly supplied by Assoc. Prof. Rosemarie Friemann
‡ N/D = Not determined as there are no aromatic residues for this construct

8.3.3 Media

8.3.3.1 Luria-Bertani media

Luria-Bertani (LB) media was prepared by reconstitution from a commercial powdered stock, containing 10 g L^{-1} tryptone, 5 g L^{-1} yeast extract and 5 g L^{-1} sodium chloride. For every 1 L of LB medium required, 20 g of powder was resuspended in MilliQ water. This media was subsequently sterilised by autoclave and stored at room temperature until further use.

8.3.3.2 Terrific Broth media

Terrific Broth (TB) media was prepared by resuspending 20 g tryptone, 24 g yeast extract, 9.4 g dipotassium phosphate, 2.2 g monopotassium phosphate, and 4 mL glycerol in 1 L of MilliQ water. Prior to be being sterilised by autoclave, the pH of the media was adjusted to 7.2 using concentrated sodium hydroxide. TB media was stored at room temperature until further use.

8.3.3.3 LB agar

For every 1 L of LB agar required, 20 g of LB powder was mixed with 12 g of agar and resuspended in MilliQ water. This media was subsequently sterilised by autoclave. To prepare LB agar plates, freshly autoclaved media was supplemented with filter sterilised appropriate antibiotic, gently swirled and poured into sterile petri dishes under aseptic conditions. Once cooled, LB agar plates were sealed with plastic wrap and stored at 4 °C for up to two to three weeks.

8.3.3.4 Super optimal broth with catabolite repression media

Super optimal broth with catabolite repression (SOC) medium was prepared by the addition of 5 g of yeast extract, 20 g tryptone, 0.5 g sodium chloride, 2.5 mL 1 M potassium chloride, 10 mL magnesium chloride and 10 mL 1 M magnesium sulfate to 900 mL of MilliQ water. The pH of the SOC media was adjusted to 7.0 with the drop-wise addition of concentrated sodium hydroxide and sterilised by autoclave. Once cooled 20 mL of a sterile 1 M glucose was added and the final volume adjusted to 1 L with sterile MilliQ water. SOC media was then aliquoted into 1.7 mL eppendorf tubes under aseptic conditions and stored at -20 °C until further use.

8.3.4 Antibiotics

Antibiotics were prepared as stock solutions and filter sterilised through a 0.22 µm syringe filter. The appropriate filter sterilised antibiotic was subsequently used to supplement autoclaved media; these are described in Table 8.3.

Table 8.3 | Antibiotic solutions used throughout this thesis. The stock concentration, final working concentration and solvent used to resuspend the antibiotic are presented.

Antibiotic	Stock concentration	Final working concentration	Solvent
Ampicillin	100 mg mL^{-1}	100 µg mL^{-1}	MilliQ water
Chloramphenicol	100 mg mL^{-1}	34 µg mL^{-1}	Ethanol
Kanamycin	50 mg mL^{-1}	50 µg mL^{-1}	MilliQ water

8.3.5 Preparation of competent cells

All strains of chemically competent *E. coli* cells used throughout this research were initially prepared and supplied commercially. Cloning and expression strains used routinely were prepared in-house using a calcium chloride-based method. Firstly, commercial stocks were streaked onto LB agar plates without antibiotic and incubated overnight at 37 °C. The following morning, a single colony was inoculated into 10 mL sterile LB medium and incubated overnight at 37 °C, without antibiotic. The following morning, 1 mL starter culture was added to 100 mL of LB and left to grow at 37 °C until an OD_{600} of 0.35-4 was reached. When the optimal OD_{600} was reached, cells were immediately put on ice and cooled for 30 min, swirling occasionally to ensure even cooling. Once cool, cells were centrifuged for 5 min at 6000 rpm. The supernatant was decanted and gently resuspended in 10 mL sterile, ice cold 0.1 M calcium chloride before a further 30 min incubation on ice. Following incubation, cells were centrifuged for 5 min at 6000 rpm. The supernatant was discarded, and cells were gently resuspended in 5 mL sterile, ice cold 0.1 M calcium chloride, 15% (v/v) glycerol. Chemically competent cells were aliquoted into sterile 1.7 mL eppendorf tubes under aseptic conditions, flash-frozen and stored at -80 °C until further use.

8.3.6 Transformation of competent cells

The heat-shock method of transformation was used to incorporate exogenous plasmid DNA into chemically competent *E. coli* cells. Competent cells were initially thawed on ice for 20 min. For each cell transformation, approximately 100 ng of plasmid DNA was added to 50 µL of competent cells and mixed

gently by stirring. Cells were incubated on ice for 20 min before being subjected to heat shock at 42 °C for 45 seconds. Cells were incubated for a further 5 min on ice. The transformed cells were recovered by adding 500 µL of SOC media and incubating cells at 37 °C for 1 hr at 250 rpm. Cells were subsequently spread onto LB agar plates with appropriate antibiotic selection and left to incubate overnight at 37 °C.

8.3.7 Preparation of glycerol stocks

A single transformed colony was selected to inoculate a 10 mL LB starter culture, supplemented with an appropriate antibiotic. Left to incubate at 37 °C overnight, an aliquot was added to a fresh 10 mL LB culture and left to grow at 37 °C, with shaking at 220 rpm, until an OD_{600} of 0.4-5 was reached. When the optimal OD_{600} was reached, an aliquot of culture was added to a sterile 1.7 mL eppendorf tube and glycerol was added to final concentration of 15% (v/v). Tubes were subsequently flash-frozen in liquid nitrogen and stored at -80 °C until further use.

8.4 Biomolecular techniques

8.4.1 Determination of protein and DNA concentrations

8.4.1.1 Bradford protein assay

Protein concentration was measured using the Bradford assay (3) where no aromatic residues were present in the sequence or when protein was in complex with DNA. To prepare a 1x dye solution, one part of concentrate Bradford dye reagent was diluted in four parts MilliQ water and left to equilibrate at room temperature. In order to construct a calibration curve, bovine serum albumin was serially diluted from 0.5 mg/mL to 0.016 mg/mL in the same buffer that the protein sample was in. Samples were prepared in triplicate by thoroughly mixing 10 µL protein with 200 µL 1x Bradford dye reagent in a 96 well microplate. A blank was prepared using 10 µL of MilliQ water in place of protein. Prior to measurement of the absorbance at 595 nm, triplicates were incubated for 5 min at room temperature. Measurements were taken using a SpectraMax M5 microplate spectrophotometer (Molecular Devices). Unknown protein samples were diluted in accordance with the midpoint of the calibration curve, where the concentration was extrapolated from the absorbance at 595 nm.

8.4.1.2 Quantitation of proteins using ultra-violet spectroscopy

The concentration of all protein and DNA in this study was determined by ultra-violet (UV) spectrometry. UV absorbance at a wavelength of 280 nm was measured using a NanoDrop™ 1000 spectrometer (Thermo Fisher). The concentration of protein was calculated using Beer-Lambert Law (Equation 8.1).

$$A = \varepsilon \, c \, l \tag{8.1}$$

Where A is the absorbance at 280 nm (A.U)

ε is the wavelength-dependent extinction coefficient ($M^{-1} cm^{-1}$)

c is the concentration of protein (mg mL^{-1})

l is the pathlength (cm)

The extinction coefficient for each protein was predicted from the sequence using ExPASy ProtParam (https://web.expasy.org/protparam/) based on the content of the aromatic residues, tryptophan and tyrosine (4). The extinction coefficients of each protein used in this research are listed in Table 8.2.

8.4.1.3 Quantitation of DNA using ultra-violet spectroscopy

The concentration of DNA was determined by the Nucleic Acid module on the NanoDrop™ 1000 spectrometer. UV absorbance at a wavelength of 260 nm was used for DNA quantitation. Furthermore, the relative purity or contamination could be assessed using the 260/280 nm absorbance ratio. A ratio of approximately 1.8 is generally accepted as 'pure' for DNA, while a ratio of approximately 0.6 is generally accepted as 'pure' for protein.

8.4.2 DNA preparation

Complementary DNA oligonucleotides (IDT) were resuspended in annealing buffer (10 mM Tris-HCl, pH 8, 100 mM NaCl), mixed at equimolar concentrations, and hybridised by heating to 95 °C for 5 min in a heat block, followed by cooling slowly to room temperature. Double stranded DNA (dsDNA) oligonucleotides were stored at -20 °C until use. All DNA oligonucleotides used in this study are listed below in Table 8.4.

Table 8.4 | DNA oligonucleotides used throughout this thesis. The sequence and application are presented. 6-carboxyfluorescein (FAM) was used as a fluorescent tag for fluorescent detection.

Construct	Sequence	Application
Two binding site model with 5'FAM (*Ec*NanR)	FAM-5'-TGGTATAACAGGTATAA-3' 3'-ACCATATTGTCCATATT-5'-FAM	EMSA/AUC
Full operator site with 5'FAM (*Ec*NanR)	FAM-5'-TTGATCTGGTATAACAGGTATAAAGGTATATCGTT-3' 3'-AACTAGACCATATTGTCCATATTTCCATATAGCAA-5'-FAM	EMSA/AUC
Two site model (*Ec*NanR)	5'-TGGTATAACAGGTATAA-3' 3'-ACCATATTGTCCATATT-5'	AUC/Cryo-EM
Full operator site (*Ec*NanR)	5'-TTGATCTGGTATAACAGGTATAAAGGTATATCGTT-3' 3'-AACTAGACCATATTGTCCATATTTCCATATAGCAA-5'	AUC/Cryo-EM
Two binding site model (sticky end) (*Ec*NanR)	5'- CTGGTATAACAGGTATAA -3' 3'- ACCATATTGTCCATATTC-5'	Crystallisation
Full operator site with 5'FAM (*Sp*NanR)	FAM-5'-TCTGAAAGTACTTTTAGA-3' 3'-AGACTTTCATGAAAATCT-5'	EMSA/AUC

8.4.3 Plasmid preparation

Plasmid DNA were purified using a DNA-Spin™ Plasmid DNA Purification Kit (iNtRon Biotechnology). For initial DNA amplification, the plasmid interest was transformed into chemically competent *E. coli* cells (DH5α strains) by heat-shock (Section 2.3.5). A single transformed colony was selected to inoculate a 10 mL LB starter culture and was incubated at 37 °C overnight. Following incubation, the cells were centrifuged, and plasmid DNA was purified according to the manufacturer's instructions. Plasmid DNA was sequence verified by next-generation sequencing (Section 8.2.2).

8.4.4 In-Fusion cloning

In-Fusion cloning is a ligation-independent cloning method and was employed in the later stages of this research as per the manufacturer's instructions (TaKaRa Bio USA). This process is explained below.

8.4.4.1 Primers

Primers for the gene of interest were designed using the online design tool (https://www.takarabio.com/). When supplied with the vector and target gene sequence the tool designs primers that maintain three characteristics: 1) 15 base-pair extension at the 5' end that is

complementary to the end of the digested vector to which it will be joined; 2) 3' end of the primer must be gene-specific, be between 20-25 base pairs in length, maintain a GC content between 40-60% and have a melting temperature between 50-65 °C; and 3) complementarity within each primer must be avoided in order to prevent hairpin structures. Once designed, these DNA oligonucleotide primers were commercially synthesised through IDT. Lyophilised primers were resuspended to a final concentration of 100 µM in nuclease-free MilliQ water (Thermo Fisher) and were stored at -20 °C until further use.

8.4.4.2 Polymerase Chain Reaction

The polymerase chain reaction (PCR) was performed using In-Fusion primers and CloneAmp HiFi PCR Premix, according to manufacturer's instructions (TaKaRa Bio USA). Briefly, the PCR reaction contained 12.5 µL of the CloneAmp HiFi PCR Premix (CloneAmp HiFi Polymerase, dNTPs and optimised buffer), 5 µL of a primer master mix (Forward and Reverse at a concentration of 5-7 pmol), 2 µL original vector (containing target gene) and 5.5 µL nuclease-free MilliQ water (Thermo Fisher). The PCR amplification was performed as per conditions in Table 8.5 using a Verti 96-well thermal cycler (Applied Biosystems).

Table 8.5 | PCR amplification reaction.

Temperature (°C)	Time (sec)	Cycles
98	30	1
98	10	30
x *	15	
72	30-60 per kb	
72	120	1
4	∞	1

* Melting temp is primer dependent (~50-65 °C)

8.4.4.3 Restriction enzyme digest

Preparation of DNA for downstream cloning is dependent upon restriction digests to cut DNA at specific recognition sites to produce linearized vector. Restriction enzymes were purchased from New England BioLabs and used to obtain an empty plasmid vector or extract a target gene with compatible ends. The components of a typical restriction double digest are shown in Table 8.6, where reactions were incubated for 3 h at 37 °C. The specific 5' and 3' restriction enzyme used to sub-clone each target gene is presented in Section 8.6.

Table 8.6 | Restriction enzyme double digest reaction. Reagents are listed for a 50 µL reaction volume.

Reagent	Amount
Plasmid DNA	Approximately 2000 ng DNA
Restriction enzyme (5')	1 µL
Restriction enzyme (3')	1 µL
10x CutSmart® buffer	5 µL
Nuclease-free MilliQ water	x µL
Total volume	50 µL

8.4.4.4 DNA gel extraction

Following restriction enzyme digest or PCR, fragmented DNA products were separated by agarose gel electrophoresis (Section 8.5.2). Once mapped, the DNA band of interest was excised from the gel using a razor blade and extracted from the gel using an agarose gel DNA extraction kit, according to the manufacturer's instructions (Roche). Purified DNA was used immediately or stored at -20 °C.

8.4.4.5 Ligation-independent recombination

The final step in the In-Fusion cloning workflow is a ligation-independent recombination reaction between the linearized target vector and the PCR product with 15 base pair extensions complementary to the vector. The components of a typical cloning reaction include: 0.5 µL 5× In-Fusion HD Enzyme Premix, 4 µL linearized vector and 2 µL PCR product in a total reaction volume of 6.5 µL. This reaction was then incubated at 50 °C for 15 min using a Verti 96-well thermal cycler (Applied Biosystems). The cloning reaction was then transformed into Stellar™ competent cells.

8.5 Electrophoresis

8.5.1 Electrophoresis buffer solutions

Buffer solutions used for either agarose gel or sodium dodecyl sulfate polyacrylamide gel electrophoresis is listed in Table 8.7.

Table 8.7 | Buffer solutions for gel electrophoresis. The solution and composition are presented.

Solution	Composition
1× Tris-acetate-EDTA (TAE) buffer	40 mM Tris-HCl pH 8.0, 0.11 % (*v*/*v*) glacial acetic acid, and 1 mM EDTA
6× DNA loading buffer	0.25% (*w*/*v*) bromophenol blue, 0.25% (*w*/*v*) xylene cyanol FF, and 30% (*v*/*v*) glycerol
4× Gel loading buffer	200 mM Tris-HCl pH 6.8, 8% (*w*/*v*) sodium dodecyl sulfate, 0.4% (*w*/*v*) bromophenol blue, and 100 mM DTT (added just prior to use)
0.5× Tris-Borate-EDTA (TBE) buffer	40 mM tris, pH 8.3, 45 mM boric acid, and 1 mM EDTA
20x Bolt™ MES SDS running buffer	50 mM 2-(*N*-morpholino)ethanesulfonic acid (MES), 50 mM Tris-HCl pH 7.3, 0.1% SDS, and 1 mM EDTA
1× Tris-Glycine SDS Running Buffer	25 mM Tris-HCl, pH 8.6, 192 mM glycine, and 0.1% SDS
SimplyBlue™ SafeStain	80 mg L^{-1} Coomassie Brilliant Blue G-250, 35 mM HCl

8.5.2 Agarose gel electrophoresis

Agarose gel electrophoresis was used to separate and visualise DNA fragments. Agarose gels were prepared by dissolving 1% (w/v) agarose in 1× Tris-acetate-EDTA (TAE) buffer, followed by heating and the addition of SYBR® Safe DNA Gel Stain to a final concentration of 1×. Once poured and set, DNA samples were mixed with 6× DNA loading buffer to a final concentration of 1× and loaded into the solidified agarose gel. For molecular mass confirmation, HyperLadder™ 1 kb was run alongside the sample DNA fragments. The agarose gel cassette was set up in a mini electrophoresis tank, submersed in 1× TAE buffer and subjected to electrophoresis for 40 min at 120 volts. DNA bands within the gel were visualised by UV transillumination.

8.5.3 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate proteins and assess the level of purity after different chromatography techniques. Sodium dodecyl sulfate is an anionic detergent that binds to the surface of polypeptides, allowing them to migrate towards the anode of an electric field. The rate of movement of the polypeptide is proportional to its respective molecular mass. Therefore, protein samples can be separated on the basis on molecular mass within the gels into discrete bands.

SDS-PAGE was performed using Bolt™ 4-12% Bis-Tris plus precast protein gels in 1× Bolt™ 2-(*N*-morpholino)ethanesulfonic acid SDS running buffer. Protein samples were prepared by mixing with 4x gel loading buffer to a final concentration of 1×, heated at 80 °C for 5 min and loaded into the precast protein gel alongside Novex™ Sharp Pre-stained Protein Standard. Electrophoresis was performed at 200 volts for 22 min and gels were stained using SimplyBlue™ SafeStain. Gels were visualised and imaged using a Bio-5000 Plus digital imaging system (Microtek).

8.6 Experimental methods

All methodology affiliated with each respective chapter is presented below except for Chapter Two and Chapter Five—this is included within the chapter as part of the manuscript. All general methodology is described above, while all constructs used in this research are presented in Table 8.2.

8.6.1 Experimental methods for Chapter Three

8.6.1.1 Cloning and expression trials of the Escherichia coli YjhB construct

The gene encoding YjhB from *E. coli* was synthesised commercially by Genscript and sub-cloned into the pWarf(-) expression vector (kindly supplied by Assoc. Prof. Rosmarie Friemann, University of Gothenburg, Sweden) using the restriction sites *Xho*I and *Bam*HI to generate pWarf(-)*yjhB*. This expression construct was then transformed into *E. coli* Lemo21(DE3) cells (New England BioLabs) and plated onto LB-Agar supplemented with kanamycin (50 μg mL^{-1}) and chloramphenicol (34 μg mL^{-1}). For the expression trials, a single colony was used to inoculate 50 mL of TB media supplemented with kanamycin (50 μg mL^{-1}) and chloramphenicol (34 μg mL^{-1}). This preculture was incubated overnight at 37 °C. The following day, 2 mL of preculture was used to inoculate 250 mL of freshly prepared TB media

along with the antibiotics, kanamycin (50 µg mL^{-1}) and chloramphenicol (34 µg mL^{-1}). Once inoculated this media was aliquoted (4 mL) into two 24 deep-well blocks (Thermo Fisher), where each row explored a different L-rhamnose (Sigma) concentration (0, 100, 250, 500, 100, and 2000 µM) and each column explored a different IPTG (Bioline) concentration (0.1, 0.4, 0.7, and 1 mM). These compounds were freshly prepared, and filter sterilised prior to addition. Initially, each L-rhamnose concentration was aliquoted individually (200 µL), then each 24 deep-well block was sealed using a breathable sealing film (Axygen) and left to grow at 37 °C with shaking at 220 rpm. To monitor the OD_{600} the remaining TB media was aliquoted into a third and fourth 24 deep-well block so the expression trial would not be disturbed. Expression was then induced at an OD_{600} ≈ 1.5-2.0 by the addition of each respective IPTG concentration (200 µL). Here, one block was transferred to 26 °C, while the other was transferred to 18 °C. In addition, the supplementary deep-well blocks were transferred to each growth temperature to serve as a non-induced sample, respectively. All blocks were left to grow overnight with shaking at 220 rpm. As a negative control, empty pWarf(-) was grown in parallel, as a target gene is required for GFP to be expressed (5). Following 14-16 h incubation, the OD_{600} of each well was measured and the cell pellet was harvested by centrifugation at 8 000 rpm for 5 min before being flash-frozen for storage at -20 °C.

8.6.1.2 Whole-cell fluorescence

Cell pellets were resuspended in 200 µL Phosphate-Buffered Saline (PBS) and left to incubate on ice for 30 min. Following incubation on ice, each resuspended sample was transferred to a Corning® 96 Well Black Polystyrene Microplate. Here, the whole cell fluorescence from the green fluorescent protein (GFP) fusion tag was measured at an excitation wavelength of 485 nm and an emission wavelength of 512 nm using a SpectraMax® M5 Microplate Spectrophotometer (Molecular Devices). To calculate the whole-cell fluorescence, the data was analysed according to the protocol in Hsieh *et al*. (2010) using Equation 8.2 and 8.3 (5). Firstly, the fluorescent readings (F) for each well were normalised (F_N) against the OD_{600} reading using Equation 8.2. Secondly, this normalised fluorescence was (F_{N1}) was divided by the normalised fluorescence of the non-induced sample at the respective temperature (F_{N2}) to account for the inherent fluorescence of *E. coli* using Equation 2.3. The final whole-cell fluorescence reading was reported in relative fluorescence units (RFU).

$$F_N = \frac{F}{OD_{600}} \tag{8.2}$$

$$\mathrm{RFU} = \frac{F_{N1}}{F_{N2}} \tag{8.3}$$

8.6.1.3 In-gel fluorescence

Following whole-cell fluorescence measurements, the top six samples with respect to RFU were sonicated on ice using a sonication bath for 30 min. Once sonicated, samples were mixed with 1× Novex® Tris-Glycine SDS Sample Buffer (Thermo Fisher), loaded into a 10-well Novex® 4-12% Tris-Glycine precast gel and subjected to electrophoresis for 1.5 hr at 125 volts using 1× Tris-Glycine SDS Running Buffer (25 mM Tris-HCl, pH 8.6, 192 mM glycine, 0.1% SDS). To provide an indication of the molecular weight for each sample, a Benchmark™ Fluorescent protein standard was run alongside. After electrophoresis, the gel was rinsed in MilliQ water and imaged with a Typhoon FLA 9500 (GE Healthcare) using a 489 nm excitation source and an LPB emission filter.

8.6.1.4 Bioinformatic analysis of E. coli YjhB

To search and compare protein sequences for *E. coli* YjhB, the basic local alignment search tool (BLAST) program and BLASTp (6) was used. The amino acid sequence for the top five search hits were sourced from the UniProt database (7). Multiple sequence alignments of YjhB homologues were generated using Clustal Omega (8). Membrane topology was predicted using the TOPCONS server (topcons.cbr.su.se/).

8.6.1.5 De novo 3D structure prediction

To construct a 3D *de novo* structure of YjhB, the PredMP server (9) was employed. Initially, when supplied an amino acid sequence, secondary structural elements were predicted using RaptorX-Property (10). Secondly, the spatial relationship between residues was predicted using RaptorX-Contact and deep transfer learning (11). Using the Crystallography & NMR System, this information was used to construct the 3D *de novo* model before being embedded into a membrane using an adapted 'Positioning of Proteins in Membranes' method.

8.6.2 Experimental methods for Chapter Four

8.6.2.1 Cloning of the Streptococcus pneumoniae NanR expression construct

The gene encoding NanR from *S. pneumoniae* was synthesised commercially in the cloning vector designated pUC57 with the restriction sites *Bam*HI and *Hind*III. In order to use a N-terminal His-tag, the *nanR* gene was digested from pUC57 using the high-fidelity restriction enzymes *Bam*HI and *Hind*III (New England Biolabs), gel purified and ligated with T4 DNA ligase (Takara) for 30 min at room temperature

to generate pET30ΔSE/*Sp*NanR_His. The pET30ΔSE expression vector, conferring kanamycin resistance was kindly supplied by Dr Hironori Suzuki. However, following an initial purification and DNA binding experiment this His-tag was not suitable. For removal, In-Fusion primers were designed (Forward-5'-AAG GAG ATA TAC ATA TGG ACA AAC CAG ATA TCG CAA CT-3', Reverse-5'-GCT CGA GCT AGG ATT TTC TTA CAC GTT TTT GT-3') to utilise the restriction enzyme sites *Nde*I and *Hind*III and ensure an ~15 base pair extension that is complementary to the pET30ΔSE expression vector. Briefly, the target gene was amplified using PCR and the CloneAmp™ DNA polymerase. Next, the expression vector was linearized with the restriction enzymes *Nde*I and *Hind*III and subjected to agarose gel electrophoresis alongside the PCR product. Once the integrity of the components was verified using DNA sequencing (Macrogen), both were purified using a DNA-Spin™ Plasmid DNA Purification Kit (iNtRon Biotechnology). Following plasmid preparation, In-Fusion cloning was then conducted according to the manufacture's protocol (Takara). This cloning reaction was then incubated for 15 min at 50 °C before being transformed into Stellar™ competent cells (Takara) to generate pET30ΔSE/*Sp*NanR. A diagnostic restriction endonuclease digestion was used to verify positive clones before the identity of the final construct was confirmed by DNA sequencing (Macrogen).

8.6.2.2 Expression and purification of S. pneumoniae NanR

The pET30ΔSE/*Sp*NanR construct was transformed into *E. coli* BL21(DE3) competent cells (Agilent) and grown in LB media with 30 mg mL^{-1} kanamycin at 37 °C and shaking at 220 rpm. Expression of *Sp*NanR was induced mid-log phase (OD_{600} ≈ 0.6) by the addition of IPTG to a final concentration of 1 mM and incubated overnight (14-16 h) at 26 °C with shaking at 220 rpm. Cells were harvested by centrifugation (Sorvall LYNX 4000 Superspeed) at 8 000 rpm for 10 min. Following centrifugation, the supernatant was discarded, and the cell pellet was flash-frozen in liquid nitrogen for storage at -20 °C until needed.

To prepare the protein for purification, the cell pellet was resuspended in lysis buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl), supplemented with cOmplete™ mini protease cocktail inhibitor (Roche), and lysed by sonication (Hielscher UP200S Ultrasonic Processor) on ice. Cell debris and insoluble material was pelleted by centrifugation at 16 000 rpm for 30 min. The resulting cell lysate was then saturated to 50 % ammonium sulphate and left to equilibrate for 1 h at 4 °C. Following precipitation, protein was pelleted at 10 000 rpm for 15 min and resuspended in 20 mM Tris-HCl, pH 8.0, 100 mM NaCl (buffer A), while the supernatant was discarded. This resuspended sample was dialysed overnight in buffer A at 4 °C. Purification of *S. pneumoniae* NanR was conducted *via* a three-step procedure: anion exchange chromatography; heparin affinity chromatography; and size-exclusion chromatography, using an ÄKTApure (GE Healthcare). The dialysed sample was applied to a HiTrap Q FF column (GE Healthcare),

pre-equilibrated in buffer A. Weakly-associated or positively-charged proteins were removed by washing the column for three column volumes of buffer A. Subsequently, the bound protein was eluted using a continuous gradient of buffer B (20 mM Tris-HCl, pH 8.0, 1 M NaCl) over 60 min. Fractions containing protein were identified by SDS-PAGE and then pooled. The pooled sample was then applied to a HiTrap Heparin HP column (GE Healthcare), pre-equilibrated in buffer A. Next, the column was washed using three column volumes of buffer A, while the bound protein was eluted using a continuous gradient of buffer B over 20 min. Following concentration to <500 µL using ultrafiltration (Sartorius), the pooled sample was applied to a Superdex 200 Increase 10/300 GL column (GE Healthcare) in buffer C (20 mM Tris–HCl, pH 8.0, 300 mM NaCl) where NanR eluted as a single peak. The absorbance at 280 nm of the purified protein was measured using a NanoDrop spectrophotometer and the concentration was estimated using a molar extinction coefficient of 27 390 M^{-1} cm^{-1} at 280 nm, as calculated by ProtParam (4). All purification steps were carried out at 4 °C. Protein that was not immediately used in experiments was flash-frozen in liquid nitrogen and stored at -80 °C.

8.6.2.3 Ligand screening using differential scanning fluorimetry

A QuantStudio 3 real-time PCR system (ThermoFisher Scientific) was used to record the thermal stability of the purified protein with and without potential substrates. Reactions consisted of 0.5 mg mL^{-1} (15 µM) protein, 5× SYPRO® Orange (prepared as a 50× stock) and 10 mM effector or DNA in a final reaction volume of 25 µL. Once prepared these were loaded into a 96-well microplate (Applied Biosystems). To prevent any evaporation during measurement an optical adhesive film (Applied Biosystems) was applied to the microplate prior to data collection. Samples were heated from 4 °C to a 95 °C at a ramp rate of 0.05 °C, taking fluorescent readings at each time point. Triplicate measurements were performed for each sample. To validate the experiment, one positive control with a well-characterised melting curve (lysozyme at 15 µM) and two negative controls (protein only and dye only) were conducted. Data analysis was performed using Protein Thermal Shift software (version 1.3–Applied Biosystems) where an apparent melting point (T_m) of each sample in °C was obtained from the lowest point of the first derivative plot.

8.6.2.4 Bioinformatic analysis with other RpiR-type sialoregulators

To compare the protein sequence of *S. pneumoniae* NanR with other RpiR-type NanR, amino acid sequences of NanR from *Clostridium perfringens*, *Haemophilus influenzae*, *Staphylococcus aureus*, and *Vibrio vulnificu*s were sourced from the UniProt database (7). A multiple sequence alignment was then generated using Clustal Omega (8). The DNA recognition site for NanR from Streptococci strains were

sourced from the RegPrecise database (12) and used to make a sequence logo *via* WebLogo (13). A homology model of *S. pneumoniae* NanR was generated using RaptorX (10).

8.6.2.5 Analytical ultracentrifugation of S. pneumoniae NanR

Sedimentation velocity experiments were performed with a XL-I Analytical Ultracentrifuge (Beckman Coulter) using 12 mm double-sector quartz cells and epon centrepieces in an An-60 Ti 4-hole rotor. Data was obtained at 42 000 rpm using 380 µL protein at a concentration of 0.5 mg mL^{-1} (15 µM) in size exclusion buffer (20 mM Tris–HCl, pH 8.0, 300 mM NaCl) and 400 µL of size exclusion buffer as a reference. A total of 300 scans were collected at 20 °C using radial absorbance scans at 280 nm and a step size of 0.003 cm. Data was fit to either a continuous sedimentation distribution [c(*s*)] or a continuous mass distribution [c(*M*)] model using *SEDFIT* (14). The buffer density, buffer viscosity and an estimate of the partial specific volume of the protein sample, based on the amino acid sequence, were determined using *SEDNTERP* (15).

8.6.2.6 Analytical ultracentrifugation of S. pneumoniae NanR in the presence of DNA

Sedimentation velocity experiments in the presence of DNA were performed with a XL-I Analytical Ultracentrifuge (Beckman Coulter) at 20 °C. Protein-DNA titrations were prepared in binding buffer (20 mM Tris–HCl, pH 8.0, 150 mM NaCl) and left to equilibrate at room temperature for 30 min before being loaded (450 µL per sector) into 12 mm double-sector quartz or sapphire cells with epon centre-pieces, and then mounted into an An-50 Ti 8-hole or An-60 Ti 4-hole rotor. Data was obtained at 50 000 rpm and 32 000 rpm in intensity mode at 495 nm. The optical density of each sample at 495 nm was 0.5 to ensure a balanced gain setting.

All data were analysed using UltraScan 4.0, release 2578 (16). Sedimentation data were evaluated by the two-dimensional spectrum analysis (2DSA) (17), which affords an unbiased hydrodynamic model, where the sedimentation coefficient and frictional ratio (f/f_0) can be floated independently. In addition, 50 Monte-Carlo iterations were performed on the 2DSA data set providing further refinement. All fitting procedures were completed using the UltraScan LIMS cluster. Each 2DSA-Monte Carlo model was visually assessed to ensure a good fit and a low RMSD using the FE Model Viewer utility in UltraScan. Peak integration of sample components was performed using the Genetic Algorithm Initialisation Utility in UltraScan. The buffer density, buffer viscosity and an estimate of the partial specific volume of the protein sample, based on the amino acid sequence, were determined using UltraScan.

In order to accurately determine the molecular weight of each species in solution, a weight-averaged partial specific volume was estimated for each protein-DNA hetero-complex using the following equation. The values used in this study are presented in Chapter Four, Table 4.2.

$$\bar{v} = \frac{M_1\bar{v}_1 + M_2\bar{v}_2}{M_1 + M_2} \quad (8.4)$$

Where M_1 is the molar mass of the protein component (Da)

M_2 is the molar mass of the DNA component (Da)

$\bar{v}_1$ is the partial specific volume of the protein (mL g^{-1})

$\bar{v}_2$ is the partial specific volume of the DNA (mL g^{-1})

8.6.2.7 Electrophoretic mobility shift assay with S. pneumoniae NanR

Double stranded 5'FAM labelled DNA oligonucleotides (IDT) were diluted to 50 nM in binding buffer (20 mM Tris–HCl pH 8.0, 150 mM NaCl, 5% Glycerol). This buffer was supplemented with 5 mM *N*-acetylmannosamine-6-phosphate to investigate the role of the phosphorylated sugar. 12-well Novex 6% Tris-Glycine gels (Invitrogen) were pre-run in 0.5× Tris-Borate-EDTA (TBE) buffer (40 mM Tris-HCl, pH 8.3, 45 mM boric acid, 1 mM EDTA) for 30 min at 200 V and 4 °C. Protein and DNA oligonucleotides were mixed and incubated at room temperature for 30 min to allow samples to reach equilibrium. Electrophoresis was performed immediately on the pre-run gels in 0.5× TBE buffer for 20 min at 200 V and 4 °C. After electrophoresis, the gel was rinsed in MilliQ water and imaged with the Typhoon FLA 9500 (GE Healthcare) using a 473 nm excitation source and an LPB emission filter.

8.6.2.8 Crystallisation and data collection of S. pneumoniae NanR

Crystallisation trials for *S. pneumoniae* NanR were performed in-house using the JCSG-*plus*, PACT-*premier*, Shotgun, Morpheus and the Clear Strategy Screens (I and II) from Molecular Dimensions. Trials were conducted using the sitting-drop vapour-diffusion method at 8 °C and 20 °C with droplets consisting of 400 nL protein solution and 400 nL reservoir solution. Purified NanR was concentrated to 10 mg mL^{-1} in 20 mM Tris-HCl, pH 8.0, 150 mM NaCl, or alternatively a buffer supplemented with 10 mM *N*-acetylmannosamine-6-phosphate. Conditions that produced crystals suitable for diffraction screening were mounted onto cryo-loops (selected based on the size of crystal), and cryo-protected by soaking in 85% reservoir solution and 15% glycerol-ethylene glycol (50:50 mix) prior to being flash-

frozen in liquid nitrogen. Once flash-frozen, these crystals were then mounted onto the MX2 Beamline at the Australian Synchrotron. Diffraction data were collected (λ = 0.9537 Å) with the EIGER × 16M detector. Data was processed and scaled using *XDS* (18) and *AIMLESS* from the *CCP4* program suite (19).

8.6.2.9 Small angle X-ray scattering of S. pneumoniae NanR

Small angle X-ray scattering (SAXS) data was collected on the SAXS/WAXS beamline equipped with a Pilatus 1M detector (170 mm × 170 mm, effective pixel size, 172 µm × 172 µm) at the Australian Synchrotron. A sample detector distance of 1600 mm was used, providing a q range of 0.006-0.5 $Å^{-1}$. Here, 50 µL of purified *S. pneumoniae* NanR at 8 mg mL^{-1} (250 µM) was injected onto an inline Superdex S200 Increase 5/150 GL SEC column (GE Healthcare), equilibrated with 20 mM Tris–HCl, pH 8.0, 300 mM NaCl, and supplemented with 0.1% (*w*/*v*) sodium azide, using a flow rate of 0.35 mL min^{-1}. To characterise the solution structure of the protein-DNA hetero-complex, protein (250 µM) and DNA oligonucleotides (125 µM) were mixed and incubated at room temperature for 30 min to allow samples to reach equilibrium. DNA was collected separately at the same concentration. Co-flow SAXS was used to minimise sample dilution and maximise signal to noise (20). Scattering data was collected in one second exposures (λ = 1.0332 Å) over a total of 500 frames, using a 1.5 mm glass capillary, at 12 °C.

Analysis of the scattering data was performed using the *ATSAS* software package (version 2.8.4)(21). 2D intensity plots were radially averaged, normalised to sample transmission, and background subtracted using *CHROMIXS*. *PrimusQT* (22) was used to perform the Guinier analysis, perform the Kratky analysis and to generate the pairwise distribution function $P(r)$, which was calculated using an indirect Fourier transform. The molecular mass of the samples was estimated using the *SAXS-MoW2* package (23), or from the Porod volume. To generate a multiphase *ab initio* reconstruction of the protein-DNA hetero-complex, 10 independent *MONSA* (24) runs were performed and then averaged using *DAMAVER* (25). Each model was further evaluated using *SUPCOMB* (26) to determine the level of consistency. The resulting averaged model was visualised in PyMOL by increasing the solvent density radius.

8.6.3 Experimental methods for Chapter Six

8.6.3.1 Thermal stability of E. coli NanR

Differential scanning fluorimetry experiments with *E. coli* NanR were performed using the Prometheus NT.48 instrument (NanoTemper Technologies). Protein was prepared at 1 mg mL^{-1} in 20 mM Tris–HCl ,pH 8.0, 150 mM. This buffer was supplemented with Neu5Ac (10 mM) or DNA (300 µM) when the thermal stability of ligands was investigated. Protein samples (10 µL) were loaded into glass capillaries

and placed into the sample holder. Detection was achieved through excitation of tryptophan residues within the protein at 280 nm, while the intrinsic fluorescence intensity was recorded at 330 nm and 350 nm. The laser intensity was adjusted to 16%, based upon the number of tryptophan residues in NanR (two per monomer). Samples were heated from 20 °C to a 95 °C at a ramp rate of 1 °C per min, taking fluorescent readings at each time point. Duplicate measurements were performed for each sample. Data analysis was performed using PR.ThermControl software (NanoTemper Technologies) where an apparent melting point (T_m) of each sample in °C was obtained by taking the first derivative of the 350/330 nm ratio.

8.6.3.2 Circular dichroism spectroscopy

A Jasco-J815 circular dichroism (CD) spectrophotometer was used to record CD spectra of purified NanR. Wavelength scans between 200 and 260 nm were performed on protein samples at a concentration of 0.05 mg mL^{-1} in a 1 mm pathlength quartz cuvette. Protein stock was diluted in 20 mM Tris-HCl, pH 8.0. Data was collected at 20 °C in 0.2 nm steps. To reduce the signal-to-noise ratio, three scans were collected and then averaged. The data was converted to mean residue ellipticity by using the optical conversion tool within the Spectral Analysis software (Jasco) by supplying the molar concentration and cell pathlength.

8.6.3.3 Electrophoretic mobility shift assay with E. coli NanR

Double stranded 5'FAM labelled DNA oligonucleotides were diluted to 10 nM in gel shift buffer (10 mM MOPS, pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, 10% glycerol). 12-well Novex 6% Tris-Glycine gels (Invitrogen) were pre-run in 0.5× TBE buffer (40 mM Tris-HCl, pH 8.3, 45 mM boric acid, 1 mM EDTA) for 30 min at 200 V, and 4 °C. Protein and DNA oligonucleotides were mixed and incubated at room temperature for 30 min to allow samples to reach equilibrium. Electrophoresis was performed immediately on the pre-run gels in 0.5× TBE buffer for 20 min at 200 V, and 4 °C. After electrophoresis, the gel was rinsed in MilliQ water and imaged with the Typhoon FLA 9500 (GE Healthcare) using a 473 nm excitation source and an LPB emission filter.

8.6.3.4 Analytical ultracentrifugation of E. coli NanR and the N-terminal DNA-binding domain in the presence of DNA

Sedimentation velocity experiments in the presence of DNA were performed with a XL-I Analytical Ultracentrifuge (Beckman Coulter) at 20 °C. Protein-DNA titrations were prepared in binding buffer (20 mM Tris–HCl, pH 8.0, 150 mM NaCl) and left to equilibrate at room temperature for 30 min before being

loaded (450 µL per sector) into 12 mm double-sector quartz or sapphire cells with epon centre-pieces, and then mounted into an An-50 Ti 8-hole or An-60 Ti 4-hole rotor. Data was obtained at both a high and low speed in intensity mode at 495 nm. The optical density of each sample at 495 nm was 0.5 to ensure a balanced gain setting.

All data were analysed using UltraScan 4.0, release 2578 (16). Sedimentation data were evaluated by the two-dimensional spectrum analysis (2DSA) (17), which affords an unbiased hydrodynamic model, where the sedimentation coefficient and frictional ratio (f/f_0) can be floated independently. In addition, 50 Monte-Carlo iterations were performed on the 2DSA data set providing further refinement. All fitting procedures were completed using the UltraScan LIMS cluster. Each 2DSA-Monte Carlo model was visually assessed to ensure a good fit and a low RMSD using the FE Model Viewer utility in UltraScan. Peak integration of sample components was performed using the Genetic Algorithm Initialisation Utility in UltraScan. The buffer density, buffer viscosity and an estimate of the partial specific volume of the protein sample, based on the amino acid sequence, were determined using UltraScan.

8.6.3.5 Crystallisation and data collection of E. coli NanR in the presence of DNA

Crystallisation trials for *E. coli* NanR were performed in-house using the Shotgun and MIDAS*plu*s screens from Molecular Dimensions. Trials were conducted using the sitting-drop vapour-diffusion method at 8 °C and 20 °C with droplets consisting of 400 nL protein solution and 400 nL reservoir solution. Purified NanR was initially concentrated to 20 mg mL^{-1} in 20 mM Tris-HCl, pH 8.0, 150 mM NaCl, mixed with DNA and left to equilibrate for 30 min on ice prior to laying screens. Trials were later adapted to include a size-exclusion chromatography step following equilibration to purify the hetero-complex from free protein and free DNA. Initially, a Superdex S-200 Increase 10/300 GL column (GE Healthcare) was pre-equilibrated in 20 mM Tris-HCl pH 8.0, 150 mM NaCl, prior to injection of the sample. Multi-wavelength detection at 220 nm, 260 nm and 280 nm was employed to identify the hetero-complex. In addition, trials with a two binding site DNA model with sticky ends (Table 8.4) were performed in parallel. Conditions that produced crystals suitable for diffraction screening were mounted onto cryo-loops (selected based on the size of crystal) and were then flash-frozen in liquid nitrogen. If required, crystals were cryo-protected by soaking in 85% reservoir solution and 15% glycerol-ethylene glycol (50:50 mix) prior. Once flash-frozen, these crystals were mounted on the MX2 Beamline at the Australian Synchrotron. Diffraction data were collected (λ = 0.9537 Å) with the EIGER × 16M detector. Data was processed and scaled using *XDS* (18) and *AIMLESS* from the *CCP4* program suite (19).

8.6.3.6 Cloning of the E. coli NanR N-terminal DNA-binding domain

The gene encoding the N-terminal DNA-binding domain of *E. coli* NanR was amplified from the pET28a/*Ec*NanR construct using In-Fusion cloning, in order to utilise an N-terminal His-tag with the restriction sites *Nde*I and *Xho*I. In-Fusion primers were designed (Forward-5'-AGG AGA TAT ACC ATG CTC TCC GAA ATG GAA GAG-3', Reverse-5'-GGT GGT GGT GCT CGA GTT AGA CGC GAG CGC GTT CG-3') to ensure a ~15 base pair extension was complementary to the pET28a expression vector. Briefly, the target gene was amplified using PCR and the CloneAmp™ DNA polymerase. Next, the expression vector was linearized with the restriction enzymes *Nde*I and *Xho*I and subjected to agarose gel electrophoresis alongside the PCR product. Once the integrity of the components was verified, both were purified using a DNA-Spin™ Plasmid DNA Purification Kit (iNtRon Biotechnology). Following plasmid preparation, In-Fusion cloning was then conducted according to the manufacturer's protocol (Takara). This cloning reaction was then incubated for 15 min at 50 °C, before being transformed into Stellar™ competent cells (Takara) to generate pET28a/*Ec*NanR-DBD (DNA-binding domain). A diagnostic restriction endonuclease digestion was used to verify positive clones before the identity of the final construct was confirmed by DNA sequencing (Macrogen).

8.6.3.7 Expression and purification of the E. coli NanR N-terminal DNA-binding domain

The pET28a/*Ec*NanR-DBD construct was transformed into *E. coli* BL21 (DE3) competent cells (Agilent) and grown in LB media, supplemented with 30 mg mL^{1} kanamycin at 37 °C and shaking at 220 rpm. Expression of *Ec*NanR-DBD was induced mid-log phase (OD_{600} ≈ 0.6) by the addition of IPTG to a final concentration of 1 mM and incubated overnight (14-16 h) at 26 °C, with shaking at 220 rpm. Cells were harvested by centrifugation (Sorvall LYNX 4000 Superspeed) at 8 000 rpm for 10 min. Following centrifugation, the supernatant was discarded, and the cell pellet was flash-frozen in liquid nitrogen for storage at -20 °C until needed.

To prepare the protein for purification, the cell pellet was resuspended in lysis buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl), supplemented with cOmplete™ mini protease cocktail inhibitor (Roche), and lysed by sonication (Hielscher UP200S Ultrasonic Processor) on ice. Cell debris and insoluble material was pelleted by centrifugation at 16 000 rpm for 30 min. Purification of *Ec*NanR-DBD was conducted *via* a two-step procedure: immobilised-metal affinity chromatography; and size-exclusion chromatography using an ÄKTApure (GE Healthcare). The dialysed sample was applied to a HisTrap FF crude column (GE Healthcare), pre-equilibrated in buffer A (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 40 mM imidazole). Weakly-associated or non-specific proteins were removed by washing the column with three column volumes of buffer A. Subsequently, bound protein was eluted using a continuous gradient of buffer B

(20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 400 mM imidazole) over 60 min. As *Ec*NanR-DBD does not contain any aromatic residues, the protein was monitored *via* the peptide backbone at a wavelength of 220 nm, while the Bradford assay (3) was used to estimate protein concentration. Fractions containing the protein of interest were identified by SDS-PAGE analysis and then pooled. Following concentration to <500 µL using ultrafiltration (Sartorius), the pooled sample was applied to a Superdex 200 Increase 10/300 GL column (GE Healthcare) in buffer C (20 mM Tris–HCl, pH 8.0, 300 mM NaCl), where *Ec*NanR-DBD eluted as a single peak. All purification steps were carried out at 4 °C. The final purity was estimated to be approximately 90-95%, with only a minor contaminant at 10 kDa. This was verified by electrospray ionisation mass spectrophotometry with molar masses of 12 863 Da and 8 607 Da, respectively. Protein that was not immediately used in experiments was flash-frozen in liquid nitrogen and stored at -80 °C.

8.6.3.8 Crystallisation of E. coli NanR N-terminal DNA-binding domain in the presence of DNA

Crystallisation trials for the *E. coli* NanR N-terminal DNA-binding domain were performed in-house using Shotgun and MIDAS-*plu*s screens from Molecular Dimensions. Trials were conducted using the sitting-drop vapour-diffusion method at 8 °C and 20 °C with droplets consisting of 400 nL protein solution and 400 nL reservoir solution. Purified protein was initially concentrated to 20 mg mL^{-1} in 20 mM Tris-HCl, pH 8.0, 150 mM NaCl, mixed with DNA and left to equilibrate for 30 min on ice prior to laying screens. Trials were performed using a two binding site DNA model with sticky ends (Table 8.4) at a protein:DNA ratio of 10:1.

8.7 References

1. DeLano, W. L. (2002). The PyMOL Molecular Graphics Program. San Carlos: Delano Scientific.
2. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry, 25*, 1605-1612.
3. Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilising the principle of protein dye binding. *Analytical Biochemistry, 72*, 248-254.
4. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research, 31*, 3784-3788.
5. Hsieh, J. M., Besserer, G. M., Madej, M. G., Bui, H. Q., Kwon, S., & Abramson, J. (2010). Bridging the gap: a GFP-based strategy for overexpression and purification of membrane proteins with intra and extracellular C-termini. *Protein Science, 19*, 868-880.
6. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*, 3389-3402.
7. UniProt, C. (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Research, 36*, D190-D195.
8. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soeding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology, 7*.

9. Wang, S., Fei, S., Wang, Z., Li, Y., Xu, J., Zhao, F., & Gao, X. (2019). PredMP: a web server for *de novo* prediction and visualization of membrane proteins. *Bioinformatics, 35*, 691-693.
10. Wang, S., Li, W., Liu, S., & Xu, J. (2016). RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research, 44*, W430-435.
11. Wang, S., Li, W., Zhang, R., Liu, S., & Xu, J. (2016). CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Research, 44*, W361-366.
12. Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., Kazanov, M. D., Riehl, W., Arkin, A. P., Dubchak, I., & Rodionov, D. A. (2013). RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics, 14*, 745.
13. Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research., 14*, 1188-1190.
14. Schuck, P., Perugini, M. A., Gonzales, N. R., Howlett, G. J., & Schubert, D. (2002). Size-Distribution Analysis of Proteins by Analytical Ultracentrifugation: Strategies and Application to Model Systems. *Biophysical Journal, 82*, 1096-1111.
15. Laue, T. M., Shah, B. D., Ridgeway, T. M., & Pelletier, S. L. (1992). *Analytical Ultracentrifugation in Biochemistry and Polymer Science*. Cambridge, The Royal Society of Chemistry.
16. Demeler, B. (2005). UltraScan - A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments. In D. J. Scott, S. E. Harding, & A. J. Rowe (Eds.), *Analytical Ultracentrifugation: Techniques and Methods* (pp. 210-230): The Royal Society of Chemistry.
17. Brookes, E., Cao, W. M., & Demeler, B. (2010). A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *European Biophysics Journal with Biophysics Letters, 39*, 405-414.
18. Kabsch, W. (2010). XDS. *Acta Crystallographica Section D: Biological Crystallography, 66*, 125-132.
19. Evans, P. R., & Murshudov, G. N. (2013). How good are my data and what is the resolution? *Acta Crystallographica Section D, Biological Crystallography, 69*, 1204-1214.
20. Kirby, N., Cowieson, N., Hawley, A. M., Mudie, S. T., McGillivray, D. J., Kusel, M., Samardzic-Boban, V., & Ryan, T. M. (2016). Improved radiation dose efficiency in solution SAXS using a sheath flow sample environment. *Acta Crystallographica Section D, Structural Biology, 72*, 1254-1266.
21. Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., Kikhney, A. G., Hajizadeh, N. R., Franklin, J. M., Jeffries, C. M., & Svergun, D. I. (2017). ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography, 50*, 1212-1225.
22. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., & Svergun, D. I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography, 36*, 1277-1282.
23. Fischer, H., Neto, M. D., Napolitano, H. B., Polikarpov, I., & Craievich, A. F. (2010). Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *Journal of Applied Crystallography, 43*, 101-109.
24. Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biochemical Journal, 76*, 2879-2886.
25. Volkov, V. V., & Svergun, D. I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. *Journal of Applied Crystallography, 36*, 860-864.
26. Kozin, M. B., & Svergun, D. I. (2001). Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography, 34*, 33-41.